DOCUMENT RESUME

ED 326 037 FL 017 780

AUTHOR	Bourret, Annie, Ed.; L'Homme, Marie-Claude, Ed.
TITLE	Les Industries de la langue: Au confluent de la linguistique et de l'informatique (The Language Utilities: At the Confluence of Linguistics and Computer Science).
INSTITUTION	Laval Univ., Quebec (Quebec). International Center for Research on Bilingualism.
REPORT NO	CIRB-K-9; ISBN-2-89219-196-3
PUB DATE	88
NOTE	144p.
AVAILABLE FROM	CIRB/ICRB, Par Casault-Universite Laval, Quebec GIK 7P4, Canada.
PUB TYPE	Information Analyses (070) -- Reports - Descriptive (141)
LANGUAGE	French
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	*Computational Linguistics; Computer Assisted Instruction; Computer Oriented Programs; *Computer Science; Computer Software Development; Courseware; Databases; Dictionaries; Discourse Analysis; Foreign Countries; *French; *Language Role; *Machine Translation; Second Language Instruction; *Speech Communication; Spelling
IDENTIFIERS	*Natural Languages

ABSTRACT

A collection of essays addresses aspects of the "Language Utilities," the general term for the area of the conjunction of computer science and linguistics. The following are English translations of the titles of the articles in the collections: "Industrialization of the French Language and Its Maintenance as an Important Language of Science and Technology"; "Evaluation of Courseware: Nature and Content of the Documentation"; "PHONO: An Applier of Categorical and Variable Phonological Rules"; "Computer Assisted Translation"; "Ten Years of METEO"; "Implantation of the Computer-Assisted Translation System"; "The Terminology Bank of Quebec: An Instrument of Social Development at the Crossroads of Technology and Linguistics"; "Termium: A Modern Tool for Communication"; "Observations on the Management of the Status of French in Computer Science"; "Measure of the Degree of Difficulty of Texts Related to Computer Science"; "Computer Science in French: Today and Tomorrow"; "The French Language: Problems of Initiation to Computer Science"; "Computer Science in French"; "Technology and Syntagm: The Case of Micro-Computer Science"; "The Center for Computer Text Analysis: A Research Laboratory Specializing in Computer Text Analysis"; "Spell Checker Developed in Quebec"; "Automatic Division of Texts into Single Words"; and "A Linguistic Module for Government Computer Science in the '90s." (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED326037

PUBLICATION
K-9

(CIRB)
Centre international de recherche
sur le bilinguisme

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
L. Laforge

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

RELAI

Recherche
en
linguistique
appliquée
à
l'informatique

LES INDUSTRIES DE LA LANGUE: AU CONFLUENT DE LA LINGUISTIQUE ET DE L'INFORMATIQUE

Édité par:
Annie Bourret
Marie-Claude L'Homme

Sous la direction de:
Pierre Auger
avec la collaboration de
Carole Verreault

1988

FL 017780

cirb
icrb

Édité par:

Annie BOURRET et Marie-Claude L'HOMME

Sous la direction de *Pierre Auger*
avec la collaboration de *Carole Verreault*,
de l'Office de la langue française

LES INDUSTRIES DE LA LANGUE :
AU CONFLUENT DE LA LINGUISTIQUE ET DE L'INFORMATIQUE

Publication K-9

1988
Centre international de recherche sur le bilinguisme
International Center for Research on Bilingualism
Québec

Le Centre international de recherche sur le bilinguisme est un organisme de recherche universitaire qui reçoit une contribution du Secrétariat d'État du Canada pour son programme de publication.

Ont contribué de façon spéciale à la publication de ce bulletin, l'Université Laval et la Compagnie IBM du Canada.

The International Center for Research on Bilingualism is a university research institution which receives a supporting grant from the Secretary of State of Canada for its publication programme.

Laval University and the IBM Company of Canada have also contributed, in a special manner, to the publication of this bulletin.

© **1988 CENTRE INTERNATIONAL DE RECHERCHE SUR LE BILINGUISME**
Tous droits réservés. Imprimé au Canada
Dépôt légal (Québec) 4ème trimestre 1988
ISBN 2-89219-196-3

TABLE DES MATIÈRES

Présentation 1
Annie Bourret et Marie-Claude L'Homme

INDUSTRIES DE LA LANGUE

L'industrialisation de la langue et son maintien comme grande langue véhiculaire de la science et de la technique 3
Pierre Auger

ENSEIGNEMENT ASSISTÉ PAR ORDINATEUR

L'évaluation des didacticiels : état et nature de la documentation 15
France H.-Lemonnier et Marie-Josée Fortin

TRAITEMENT DE LA PAROLE

PHONO : un applicateur de règles phonologiques catégoriques et variables 27
Claude Paradis

TRADUCTION AUTOMATIQUE ET TRADUCTION ASSISTÉE PAR ORDINATEUR

La traduction assistée par ordinateur (TAO) 37
Anne Bordé

Dix ans de MÉTÉO 43
John Chandioux

Implantation du système de traduction assistée par ordinateur 47
Moïse Lévy

BANQUES DE TERMINOLOGIE

La Banque de terminologie du Québec : un instrument de développement social, au carrefour des enjeux technologiques et linguistiques 51
Jean-Marie Fortin

TERMIUM : outil moderne de communication 57
Winston G. Wells

INFORMATIQUE EN FRANÇAIS

Remarques sur l'aménagement du statut du français en informatique 61
Jean-Claude Boulanger

Mesure du degré de difficulté de textes relatifs à l'informatique 71
Martine Bourque

L'informatique en français : aujourd'hui et demain 77
Pierre-Étienne Laporte

La langue française: problèmes d'initiation à l'informatique 87
Michaël Mepham

L'informatique en français 95
Conrad Ouellon

Technolecte et syntagme : le cas de la micro-informatique 103
Jean Quirion

TRAITEMENT DES LANGUES NATURELLES

Le Centre d'Analyse de Textes par Ordinateur : un laboratoire de recherche spécialisé dans l'analyse de textes par ordinateur 109
Jules Duchastel

Correcteur orthographique développé au Québec 119
Richard Kromp

Découpage automatique de textes en mots simples 123
Jacques Ladouceur

Un module linguistique pour l'information gouvernementale des années 1990 141
Richard Parent

PRÉSENTATION

Annie Bourret et Marie-Claude L'Homme

Présentation

Dans le sens de sa nouvelle orientation vers l'aménagement linguistique priorisant les industries de la langue (IDLL) et fidèle à son travail de pionnier dans le domaine de la linguistique informatique, le CIRB a réuni dans ce recueil une série de textes, afin de donner un aperçu de la problématique dans ce domaine, des acquis et des défis qui restent à relever.

Les industries de la langue constituent un domaine encore peu connu dont le concept, né dans un contexte francophone, a émergé en 1984 et a commencé à être diffusé en 1985. L'ensemble des pays francophones, lors des travaux de la Conférence des chefs d'État et de Gouvernement des pays ayant en commun l'usage du français[1], ont inscrit les industries de la langue dans leur programme de préoccupations majeures. Depuis, le terme *industries de la langue* a été véhiculé et apprêté au goût des auteurs qui l'ont utilisé pour faire état de préoccupations diverses : la langue française dans la diffusion des techniques et des sciences, l'informatique en français, la linguistique informatique, ...

Lieu de conjonction de la technologie informatique ainsi que de la recherche et de l'étude systématique des langues, les IDLL désignent de façon générique une série d'applications informatiques dont la conception ne peut se faire sans apport, à un moment donné, de connaissances linguistiques. Les IDLL mettent l'ordinateur au service de la linguistique mais, de plus, mettent la langue au service de l'ordinateur : elles contribuent à la recherche en linguistique et offrent des produits informatiques commercialisables. Par ailleurs, les IDLL sont apparues comme une solution de sauvegarde de la langue française en tant que langue véhiculaire de la science et de la technique. En offrant des produits comme des algorithmes d'analyse du langage naturel, des systèmes de traduction automatique, la langue française pourrait se ranger parmi les grandes langues diffusantes dont fait partie, évidemment, l'anglais. Les applications des IDLL trouvent usage dans des secteurs aussi variés que ceux de l'éducation, de la santé, de la gestion et de la sécurité.

Malgré les nombreux acquis - notamment en terminologie, en traductique et en aide à la rédaction - il reste de nombreux problèmes, qu'un aménagement linguistique de l'informatique plus judicieux devrait permettre de résoudre. Nombre d'auteurs soulèvent les difficultés à obtenir un matériel décent utilisable en français. Les IDLL, en plus d'appartenir à un domaine hautement technique, se chargent d'une connotation affective lorsqu'il s'agit des enjeux même de la production de cette technologie.

■■■

[1] Conférence tenue à Paris du 17 au 19 février 1986.

Plusieurs des textes réunis dans ce recueil ont fait l'objet d'une communication lors de la Deuxième Conférence des chefs d'Etat et de Gouvernement des pays ayant en commun l'usage du français[2] et témoignent de la volonté des pays francophones de poursuivre les travaux dans le domaine des IDLL.

Les textes donnent un aperçu général des domaines que regroupent les industries de la langue (quoiqu'il s'agisse dans la plupart des cas d'applications spécifiques). Les IDLL restent encore difficiles à circonscrire et malgré leur apparente nouveauté, elles rassemblent des sphères d'activité moins récentes.

Les IDLL englobent :

1. le traitement des langues naturelles : l'analyse, la compréhension et la génération automatiques du langage naturel;
2. le traitement de la parole : reconnaissance et synthèse vocales, analyse du signal;
3. la traduction assistée par ordinateur et la traduction automatique;
4. la fabrication de produits informatiques : logiciels de langue française;
5. la terminotique et l'informatique lexicographique : aides aux tâches lexicographiques et terminologiques, nouveaux médias de diffusion;
6. les aides à la rédaction : traitement de texte avancé ou non, éditique, correcteurs variés;
7. la reconnaissance automatique de caractères;
8. l'informatique documentaire.

Il n'a pas été possible dans le cadre de cette publication de couvrir l'ensemble des domaines des industries de la langue et des axes de recherche de chacun d'entre eux. Mis à part le premier texte qui porte sur le thème général des industries de la langue, les textes abordent un aspect particulier du domaine. Ceux-ci touchent de l'enseignement assisté par ordinateur (évaluation des didacticiels), du traitement automatique de la parole (description d'un logiciel d'application de règles phonologiques), de la traductique, du traitement des langues naturelles (analyse de textes en langue naturelle, correcteur orthographique, lemmatisation), de la terminologie (banques de terminologie, étude sur le vocabulaire); les textes traitent également des limites et des problèmes de l'informatique en français, tant sur le plan technique que sur le plan politique ou social (aménagement linguistique en informatique, degré de difficulté des textes relatifs à l'informatique, rôle des organismes d'État).

[2]La conférence s'est tenue à Québec les 3 et 4 septembre 1987, la journée réservée aux industries de la langue a eu lieu le 5 septembre 1987.

L'INDUSTRIALISATION DE LA LANGUE FRANÇAISE ET SON MAINTIEN COMME GRANDE LANGUE VÉHICULAIRE DE LA SCIENCE ET DE LA TECHNIQUE[1]

Pierre Auger
Langues et linguistique
Université Laval

On parle beaucoup, ces dernières années, de la situation de la langue française comme grande langue véhiculaire à l'échelle planétaire. Ces bilans, presque toujours négatifs, laissent entendre que le français est en perte de vitesse et que la francophonie, telle une peau de chagrin, est en train de rétrécir dangereusement ses zones d'influence. Les uns blâment avant tout une démographie déclinante, caractéristique des pays industrialisés du Nord, d'autres l'influence dominante de l'anglo-américain sur tout le domaine scientifique et technique qui accompagne le pouvoir économique. Les plus idéalistes s'en prennent à l'incurie des gouvernements et des pouvoirs publics et leur politique défaitiste du laisser-faire. Enfin, les plus blasés veulent bien croire et répandre que l'époque de gloire du français est passée, qu'il faut être de son temps et accepter cette domination linguistique comme une condition sine qua non au progrès engendré par la modernité. Quoiqu'il en soit, malgré la divergence des points de vue, tous sont d'accord pour dire que le français est en train de vivre la pire crise de croissance de sa longue histoire, une saga autodestructrice en quelque sorte.

Cette question délicate sera appréhendée sous l'angle du concept de langue véhiculaire tel que défini par le sociolinguiste français Louis-Jean Calvet, c'est-à-dire «une langue utilisée pour l'intercommunication dans des situations plurilingues» (Calvet 1981:71-91). Dire du français qu'il doit demeurer une langue véhiculaire, c'est en quelque sorte affirmer en même temps qu'il doit continuer d'appartenir au groupe très réduit (sélect, faudrait-il écrire) des langues d'intercommunication sur la scène internationale qui trans- cendent en quelque sorte la «difficulté de communiquer» qui est le lot des langues ou groupes de langues vues comme trop étrangères ou exotiques pour servir d'interlingua dans les échanges internationaux. Selon Calvet, les facteurs dominants qui amènent une langue à jouer ce rôle sont d'ordre géographique, économique, politique, religieux et historique (prestige) principalement (Calvet 1981:77-95). Dans cette concurrence que le français doit vivre avec l'anglais américain dans son statut de langue véhiculaire, il faut isoler particulièrement les facteurs économique et politique; nous le verrons en profondeur plus loin. Les pays riches sont puissants et possèdent des langues fortes et inversement. Le poids linguistique de ces langues va d'ailleurs continuer à s'accroître au même rythme que leur richesse et leur pouvoir. Si l'on applique ce principe au rôle véhiculaire que joue une langue dans le domaine de la science et de la technique, il y a fort à parier que le pays qui produit le volume le plus important d'innovation en ces domaines verra sa langue jouer ce rôle presque sans concurrence. C'est ce qui est arrivé à la langue américaine et, sans nul doute, rien ni personne ne pourront y changer quoi que ce soit, il faut bien se le dire.

C'est ainsi que le statut de l'anglo-américain comme langue véhiculaire a continué de se consolider et de progresser avec le développement fulgurant de l'informatique durant la décennie actuelle et l'informatisation en profondeur de la société qui a suivi. Les Américains ont été les premiers et, conséquemment, ils seront sûrement les derniers à voir s'atrophier leur espace linguistique international. Cette informatisation, il y a à peine dix ans, limitée essentiellement à certains secteurs ré-

[1]Communication présentée lors du 56e Congrès de l'ACFAS, Moncton, 9-13 mai 1988.

servés des entreprises ou des centres de recherche et dirigée par des gourous informaticiens tout-puissants a pris un essor démesuré, socialement parlant, avec l'arrivée de la micro-informatique et l'envahissement de toutes les zones de la société et de la vie dite privée (famille, loisirs, etc.). On chercherait en vain un facteur d'implantation linguistique plus puissant que cette informatisation galopante qui a joué en faveur de l'anglais américain ces dix dernières années. Il faut vivre désormais avec cette idée qu'il est illusoire de penser que le français pourrait voir renversée à son profit cette situation, tout au plus peut-il aspirer à conserver une deuxième place si l'incurie ne gagne pas ses locuteurs, qu'ils soient politiciens, chercheurs, spécialistes, industriels ou simples usagers. Le chercheur Richard Parent, familier de ces questions, fait un bilan bien réaliste de la situation lorsqu'il écrit : «Il est donc important, en considérant le marché francophone de reconnaître le fait de la domination de l'anglais par l'ampleur et l'excellence de la recherche développement qui se fait dans cette langue. Du point de vue des usagers francophones de l'informatique, une version bien traduite et adaptée, ou facile à adapter au français est une nécessité. Mais il y a des différences culturelles, linguistiques, sociales, liées aux organisations qui nécessitent des adaptations plus délicates devant s'appuyer sur la recherche ergonomique, anthropologique, psychologique, linguistique, etc.» (Parent 1986:4). C'est dans cette voie que nous poursuivrons notre exposé et que nous tenterons de démontrer que si la cause est entendue, tout espoir n'est pas perdu pour l'expansion de la langue française dans le domaine crucial de l'informatique moderne. Un défi important est à notre portée et, en tant que francophones, nous devrons le relever.

Ce défi ne doit pas se définir comme une lutte à l'anglo-américain comme nous venons de le voir. Il s'agit moins de lutter contre un envahisseur potentiel ou réel, mais de préserver sa place dans ce très petit club des sociétés ayant la chance de parler une langue qui est en même temps une langue véhiculaire de transmission du savoir sur le plan international. Cela ne devrait rien enlever à l'importance de vouloir relever le défi. Un autre phénomène vient ajouter tout son poids à cette question et là, le français est solidaire des autres grandes langues du monde. En effet, l'informatisation profonde que la plupart des États modernes vivent a des retombées culturelles négatives qui ne sont pas à négliger. Nous savons tous le rôle dominant que joue encore l'écrit dans nos sociétés et, facteur nouveau, ce sont les ordinateurs qui gèrent la plus grande part de la masse des écrits que produisent nos sociétés. Cette gestion massive de l'écrit par l'informatique n'a en fait rien d'anodin pour les langues traitées, et là, le français rejoint le lot des autres grandes langues modernes des pays industrialisés utilisant une informatique d'abord conçue pour l'anglais : - logiciels de traitement de texte supportant difficilement les caractéristiques graphiques des langues nationales, - claviers mal adaptés, - logiciels de correction déficients, - documentation traduite, etc. L'ancien ministre français de la Culture Jack Lang faisait une intervention qui allait dans ce sens, en novembre 1986, au Colloque de Tours[2] : «L'incapacité d'un automate à gérer les particularismes linguistiques (par exemple, les accentuations en français, les écritures non latines) influe gravement sur le média et donc sur le contenu. L'incapacité de ce même automate à réaliser pour une langue ce qu'il peut faire pour une autre (par exemple, une aide à la correction orthographique ou à la rédaction) met la première en situation d'être supplantée par la seconde, simplement parce qu'il sera moins coûteux de produire dans une langue pour laquelle on disposera des machines les plus performantes» (Lang 1986:17-18). Cette situation décrite par M. Lang évoque bien l'enjeu culturel qui accompagne l'actuelle mutation technologique pour les peuples. C'est un peu dans les mêmes termes que le Français Bernard Cassen décrivait la portée de cet enjeu lors du même colloque : «Mais cet enjeu, et il est pourtant de taille, est en passe d'être supplanté par un autre, décisif celui-là : certains des outils les plus puissants de la recherche - les ordinateurs - sont de nature à déstabiliser complètement notre langue, en la mettant hors jeu dans la communication homme-machine, puis dans l'ensemble de la communication écrite, déjà presque totalement tributaire de l'ordinateur, et en générant des produits qui, graduellement, envahiront l'univers quotidien du simple citoyen» (Cassen 1986:13). A moyen et à long terme, ce sont donc les langues capables d'être traitées par les ordinateurs et de converser avec ces derniers qui domineront le marché de l'écrit entendu dans son sens économique, culturel et linguistique. Au rythme où vont les choses peu d'entre elles seront encore au rendez-vous en l'an 2000.

---

[2] "Les industries de la langue. Enjeux pour l'Europe", Tours, 28 février - 1er mars 1986.

Il est important ici d'ouvrir une parenthèse pour introduire un concept nouveau qui a fait son apparition à partir de 1985 à peu près et qui connaît une carrière florissante depuis lors : les «Industries de la langue». Une présentation de ce concept relativement nouveau et son application au développement du français va situer le débat qui nous occupe aujourd'hui vers une perspective beaucoup plus large qui nous fera dépasser la simple question «Faut-il franciser l'informatique?» pour en poser une nouvelle, plus décisive «Peut-on se passer de l'informatique en français pour survivre en tant que francophones?» On parle de plus en plus ces dernières années de ce concept nouveau en linguistique appliquée. Les «Industries de la langue» sont à rattacher directement à tout le domaine de l'informatique d'orientation textuelle. Ce concept a été défini lors du Premier sommet de la francophonie de Paris[3] de la façon suivante «[Les industries de la langue sont celles qui] fabriquent et commercialisent des automates qui manipulent interprètent, génèrent le langage humain, aussi bien sous sa forme écrite que sous sa forme parlée» (*Rapport de synthèse : Industries de la langue* 1986:86). De façon plus précise, on peut encore définir les industries de la langue comme «des activités de développement, de production et de commercialisation des nouvelles technologies de l'information (NTI) qui font appel à la fois à l'informatique (ordinateurs et logiciels) et aux résultats de l'étude systématique des langues. Elles développent des produits (machines) capables de traiter des informations linguistiques et susceptibles de communiquer ces informations entre eux et également avec les humains» (*Rapport de synthèse : Industries de la langue* 1986:86). Une définition plus pragmatique nous est donnée par William Baranes dans la revue *Qui-vive international* lorsqu'il écrit que «(Or) cette manipulation «machinique» nécessite des produits industriels. On assiste donc à la naissance de véritables «industries de la langue», regroupant d'une part les artisanats traditionnels du langage (traduction, dictionnaires, etc.) qui s'industrialisent grâce aux moyens informatiques, et, sous la poussée de la demande, d'autre part, un ensemble d'activités nouvelles, au confluent de l'informatique et de la linguistique (traitement automatique du langage naturel, synthèse et reconnaissance de la parole, etc.) (Baranes 1986:74).

Ces produits issus des industries de la langue peuvent être classés en trois grandes catégories (*Rapport du Sous-comité des Industries de la langue au Comité de coordination des contributions québécoises* 1987) :

a) Les outils de développement de la langue

Il s'agit essentiellement d'outils internes de recherche et de développement qui sont mis à la disposition des linguistes, des terminologues et des traducteurs, et qui sont nécessaires à la modernisation de la langue et à sa nécessaire adaptation à l'évolution technologique. On peut classer dans cette catégorie les travaux de recherche linguistique qui comprennent tous les instruments physiques, conceptuels (algorithmes) ou les logiciels utiles aux recherches terminologiques, linguistiques et sémantiques : analyse de texte, reconnaissance de termes complexes, indexation, modélisation de processus, etc. On peut rattacher à ce groupe les réseaux de terminologie et les banques de données linguistiques et terminologiques qui sont d'ores et déjà des outils très performants d'accès à l'information de type linguistique.

b) Les outils d'utilisation de la connaissance linguistique

On situe dans cette catégorie les outils qui permettent de traduire dans des applications de la vie courante les résultats de la recherche dans les différentes sphères de l'informatique linguistique. Les systèmes de traduction automatique sont à classer dans cette seconde catégorie. Ils constituent

[3]Conférence des chefs d'État et de Gouvernement des pays ayant en commun l'usage du français, Paris, 17-19 février 1986.

un des domaines les plus anciens de la recherche en intelligence artificielle et ce n'est que maintenant, tout juste après trente ans de recherche, que l'on commence à obtenir des résultats intéressants dans ce domaine. On peut encore citer, dans la même catégorie, les banques de données textuelles, les systèmes d'interprétation du langage naturel, les systèmes de contrôle vocal pour les machines, les systèmes experts et les didacticiels.

c) Les outils de réalisation de produits linguistiques

La quasi totalité des textes produits aujourd'hui transitent, à un moment donné de leur existence, par un ordinateur. C'est donc dire l'importance que prend aujourd'hui l'informatique à orientation textuelle. Pour une langue comme le français, ce fait est lourd de conséquences. En effet, comment concevoir l'avenir du français comme langue véhiculaire dans un environnement informatique presque exclusivement dominé par la langue anglaise? La langue de «travail» des ordinateurs est presque exclusivement l'anglais et les logiciels sont le plus souvent conçus en fonction de l'anglais (ex. problèmes des claviers, classement des unités linguistiques par les **s.g.b.d.**[4] sans prise en compte des signes diacritiques du français, absence de reconnaissance des structures syntaxiques du français, etc.). Il va sans dire que l'avenir du français comme langue véhiculaire de la science et de la technique est aussi lié au développement de ces outils orientés vers le texte. Parmi ces produits qui visent la réalisation de produits linguistiques ou langagiers (le texte, sous toutes ses formes), mentionnons les logiciels de traitement de texte, les progiciels grammaticaux ou les correcteurs orthographiques, les logiciels d'éditique, les logiciels d'indexation ou de génération de textes, les dictionnaires et les thésaurus sur disque compact (CD-ROM) pour ne mentionner que les produits les plus courants. On peut donc voir l'importance des IDLL[5] comme domaine de recherche dans le monde contemporain, domaine carrefour pour tout le progrès concernant les sciences de l'information. Lors du colloque de Tours[6], Bernard Cassen, un des précurseurs visionnaires en ce domaine, affirmait avec force que «(On peut même dire) que les langues qui ne s'industrialiseront pas cesseront, à un terme plus ou moins bref, d'être véhiculaires, parfois même dans d'importants secteurs d'activité des pays où elles sont parlées» (Cassen 1986:13). Plus récemment, le linguiste Claude Hagège écrivait dans le même sens : «Une attention toute particulière doit également être accordée à ce que l'on appelle les industries de la langue, c'est-à-dire l'application industrielle du traitement de la langue, en tant que matériau, par les machines informatiques, afin de répondre à un certain nombre de besoins de secteurs tels que l'éducation, la santé, la sécurité, etc. Si le français ne s'industrialise pas, ses chances de se maintenir au niveau d'une langue internationale iront en s'amenuisant, du fait de la pression croissante qu'exerceront, dans tout type de communication industrielle, commerciale et scientifique, les ordinateurs capables de manipuler l'anglais, c'est-à-dire la langue des ingénieurs qui construisent actuellement le plus d'ordinateurs» (Hagège 1987:249). C'est à la même conclusion qu'arrivent les chefs d'État des pays francophones réunis à Québec au mois de septembre 1987[7]; on peut lire, en effet, dans les documents de conférence publiés lors du Sommet de Québec que : «La langue française doit s'inscrire rapidement dans le mouvement actuel d'industrialisation des langues, autrement elle deviendra de moins en moins apte au développement de la recherche et de la production dans les secteurs de pointe et, à long terme, se marginalisera par rapport aux autres grandes langues de communication internationales dans ces champs d'activité essentiels à l'avenir de la francophonie» (*Documents de conférence* 1987:173). Il

[4] S.g.b.d. = système de gestion de bases de données.

[5] IDLL est un acronyme pour Industries de la langue.

[6] "Les industries de la langue. Enjeux pour l'Europe", Tours, 28 février - 1er mars 1986.

[7] Deuxième Conférence des chefs d'État et de Gouvernement des pays ayant en commun l'usage du français, Québec, septembre 1987.

faut noter que sous ces discours, leurs auteurs laissent entendre que les efforts actuels des pays francophones pour industrialiser la langue française ne sont pas suffisants et qu'il y aurait urgence de remédier à cette situation.

Dans le cadre de la coopération franco-québécoise et de la préparation de la future structure des Sommets francophones, les gouvernements français et québécois mettaient sur pied en 1985 un groupe de travail bilatéral dit «Groupe de réflexion et d'action sur les enjeux de la francophonie (Harvey de Saint-Robert)». Les résultats contenus dans le rapport du Groupe de travail identifiaient clairement des besoins de développement de l'informatique en langue française (domaine que l'on recouvrira bientôt sous le générique «industries de la langue»). Ce premier constat d'un groupe francophone identifiait prioritairement «l'urgence de la création en langue française de logiciels originaux et de l'adaptation de cette même langue aux produits existants. Un espace culturel et économique francophone, souhaité par plusieurs, ne pourra évidemment pas se constituer sans qu'un effort concerté et important ne soit consenti à l'édification dans nos pays d'une industrie du logiciel, fer de lance efficace et novateur de l'avènement de la société de demain, celle de l'information et des communications» (*Actes de la Conférence* 1986:45). Dans cette foulée, le Groupe identifiait des besoins prioritaires quant à la diffusion de l'information sur le stock de logiciels disponibles en langue française et à la conception, la production et la distribution des logiciels en langue française. Pour ce deuxième aspect, le Groupe de travail identifiait plusieurs éléments touchant l'industrie du logiciel en langue française, comme l'utilisation dans le commerce et l'industrie de logiciels en langue française pour accompagner le choix du français comme langue de\du travail, l'adaptation de logiciels originellement en langue anglaise pour le bénéfice de tous les francophones (coopérations multilatérales), formation des utilisateurs, des concepteurs de logiciels, des techniciens ..., élaboration de didacticiels qui ont un poids linguistique important et développement d'une terminologie informatique en langue française unifiée en concertation avec toutes les régions de la francophonie. Le même Groupe proposait en guise de conclusion un train de mesures touchant quatre volets principaux (*Actes de la Conférence* 1986:44-45) :

1- constitution d'une banque de données sur les logiciels disponibles en langue française;
2- mise sur pied de programmes de formation et de coopération avec l'ACCT touchant l'informatique en langue française;
3- élaboration d'un dictionnaire encyclopédique de l'informatique en langue française et enfin,
4- établissement de structures favorisant des ententes commerciales et industrielles multilatérales entre producteurs de logiciels.

Nous verrons que ces orientations se sont modifiées quelque peu dans les années suivantes avec les travaux du Réseau des industries de la langue, créé dans la structure des Sommets francophones. Disons enfin que le Groupe remettait aux gouvernements des États francophones des responsabilités importantes d'animation, de financement et de contrôle de l'espace informatique francophone.

Vers la même époque, un autre groupe de travail, Canada- Québec cette fois, a été mis sur pied par les ministères des communications canadien et québécois pour étudier la question du développement du logiciel d'expression française. Dans le rapport qu'il a rendu public en 1986 touchant cette question, le Comité décrit dans le détail les caractéristiques du marché québécois du logiciel, les contraintes de l'industrie évoluant sur le continent américain, les besoins de francisation qui s'expriment au niveau du travail et du commerce et enfin, le rôle que devrait jouer l'État en ce domaine (*Rapport d'étape du Comité Canada-Québec* 1986). Fait intéressant, le Comité a recueilli plusieurs éléments sur les caractéristiques (tendances) linguistiques des usagers industriels québécois (appréhensions, choix du standard linguistique, acceptabilité langagière etc.). Voici quelques-uns des problèmes relevés dans le rapport du Comité et reliés à la francisation informatique :

1- les logiciels en français ne représentent pas un standard (ils ne fonctionnent pas avec toutes les imprimantes, ils nécessitent des ajustements parfois très laborieux sur les claviers, etc.);

2- les logiciels américains les plus courants semblent gêner à un degré moindre les utilisateurs francophones dans la mesure où ils font appel à un vocabulaire limité de nature mnémonique;
3- la langue constitue une variable particulièrement importante pour certains types de logiciels comme le traitement de texte, la comptabilité et les produits destinés à un usage domestique;
4- les logiciels (ainsi que la documentation) conçus en France soulèvent parfois des incompréhensions quant à la terminologie utilisée et des incompatibilités quant aux pratiques informatiques;
5- les acheteurs craignent que la version française des produits de langue anglaise ne soit pas la plus récente, ou que les versions évoluées du produit ne soient plus disponibles en français.

Les résultats obtenus par le Comité mixte Canada-Québec mettent ici en évidence la nécessité d'une approche très adaptée à la situation économico-linguistique du Québec et de l'Amérique, difficilement conciliable, peut-être, avec l'approche globalisante et panfrancophone suggérée dans les travaux des Sommets francophones depuis 1986. Pour l'industrie en général, le fait d'appartenir à une Amérique majoritairement anglophone impose des contraintes dont il faut tenir compte dans un contexte de francisation informatique. Malgré les arguments démontrant que la disponibilité de logiciels en langue française n'est pas toujours une réalité, que les coûts de francisation sont trop élevés pour l'accès à un marché aussi étroit, le Comité a retenu une pénurie réelle de logiciels en langue française dans un certain nombre de domaines clés comme les logiciels comptables, les tableurs et les bases de données. La situation est encore plus critique dans les domaines pointus des sciences, du génie et des systèmes experts. Le Comité note encore que : «La demande s'oriente de plus en plus vers les produits normalisés et validés (progiciels), paramétrés (offrant la flexibilité nécessaire pour adapter certaines caractéristiques, la langue d'affichage par exemple, aux besoins de l'usager) et «conviviaux» (user-friendly)» (*Rapport d'étape du Comité Canada-Québec* 1986).

Quant à la traduction, le Comité note qu'elle doit s'effectuer en respectant la culture du milieu auquel elle est destinée: « ... il importe, lors de la traduction d'un logiciel en d'autres langues, de s'adjoindre des traductions qui ont un lien direct avec la culture actuelle du pays auquel est destiné le logiciel, et de préférence avec une communauté nationale familière avec la sphère d'activité dans laquelle s'inscrit le logiciel.» (*Rapport d'étape du Comité Canada-Québec* 1986). Il est également suggéré qu'une attention particulière soit portée au choix de la terminologie apparaissant dans les menus de logiciels et dans les documents les accompagnant : « ... La terminologie utilisée, tant dans la conception d'un logiciel que dans sa traduction ou son adaptation pour un pays, doit refléter une langue internationale. Les traductions de documentation américaine effectuées parallèlement ici et en Europe ont déjà fait naître des néologismes aux deux endroits qui risquent de créer des divergences d'habitudes, difficiles à corriger, qui nuisent à la communication internationale.» (*Rapport d'étape du Comité Canada-Québec*, 1986). Quant à la terminologie utilisée dans les logiciels produits en France, le Comité note: «Les logiciels (ainsi que la documentation) conçus en France soulèvent parfois des incompréhensions quant à la terminologie utilisée et des incompatibilités quant aux pratiques informatiques» (*Rapport d'étape du Comité Canada-Québec* 1986). Nous discuterons plus loin cette assertion.

Conscient de la situation hautement complexe où se situe l'informatique et son utilisation au Québec, l'Office de la langue française réunissait un colloque à l'automne 1986[8]. Cet événement organisé conjointement par l'Office et la Fédération de l'informatique du Québec visait à démontrer la disponibilité de la technologie informatique en français au Québec et la rentabilité économique et sociale qui allait en découler rapidement. Les conférenciers invités ont mis en évidence un certain nombre de faits «stratégiques» concernant l'industrie québécoise de l'informatique et son développement concurrentiel face au marché. Le président de l'Office d'alors, M. Claude Aubin, affirmait à cette occasion que : «La démocratisation de l'informatique entraînant une diminution des coûts, l'ac-

[8]Colloque "Réussir l'informatique en français", Montréal, automne 1986.

cessibilité accrue aux logiciels ainsi que la convivialité que ceux-ci ont atteint, constituent des assises fermes reposant sur les acquis propres à la communauté francophone. Il faut se souvenir qu'en tant que consommateurs francophones de l'informatique, nous constituons une clientèle qui a des droits à faire valoir en tant que collectivité attachée à sa langue» (*Francisation en marche* 1986).

Il n'en demeure pas moins que malgré ces propos qui se veulent rassurants un certain nombre de problèmes sont demeurés sans solution ou partiellement solutionnés. M. Conrad Ouellon, organisateur de ce Colloque qui nous réunit aujourd'hui, dans une conférence qu'il a donnée à Québec ce printemps sur l'informatique en français, identifiait trois zones grises qui causent le plus de problèmes à l'utilisateur québécois (Ouellon 1986:9-11):

1- le matériel avec des claviers diversement francisés mais toujours d'utilisation malaisée (manipulation des signes diacritiques du français, touches identifiées en anglais, etc.) et des imprimantes réticentes à reproduire esthétiquement certaines particularités du français comme les majuscules accentuées, matériel qu'il faudra, de toute façon, veiller à configurer péniblement et parfois même en vain. La situation est pire encore avec les lecteurs optiques utilisés pour le texte;
2- les logiciels, le plus souvent en anglais ou traduits, mais ne reconnaissant guère le français quand il s'agit d'effectuer des tris ou certaines opérations (cf. le traitement que réserve **DBASE III+** au «è» lorsqu'on ajoute des dossiers à un fichier *.dbf et son incapacité à indexer des mots accentués sans artifice). Il en est de même en général des analyseurs de texte, des lemmatiseurs, des logiciels d'indexation et de reconnaissance de texte conçus d'abord pour l'anglais;
3- les manuels souvent mal traduits ou incompréhensibles même quand ils sont conçus originellement en français: multiplicité des synonymes, utilisation de termes ambigus ou non définis et calques, auxquels problèmes s'ajoute l'absence d'une terminologie informatique unifiée ou internationale, nous reviendrons plus tard sur ce point.

Le conférencier offrait à cette occasion des solutions à court et à moyen termes comme l'élaboration sur une base internationale de «standards adaptés» au traitement des autres langues que l'anglais pour les problèmes de matériel, la conception de logiciels modulaires multilingues permettant une meilleure intégration linguistique des produits informatiques pour éviter les hiatus culturels amenés par la simple traduction des logiciels et enfin, la rédaction de manuels d'utilisation en une langue claire et compréhensible pour l'usager, ce qui n'est pas un objectif simple à atteindre, nous le verrons (Ouellon 1986:13).

Le Québec n'est pas le seul État francophone à vivre des problèmes de convivialité avec l'informatique telle qu'implantée par les Américains à l'échelle planétaire, les francophones européens, en particulier, doivent vivre des problèmes analogues à ceux qu'a mentionnés M. Ouellon. Si l'on se déplace cette fois vers le développement informatique en langue française, les problèmes «superficiels» que nous venons d'évoquer paraissent bien ténus à côté de ceux rencontrés dans la recherche informatique elle-même, l'utilisation des nouveaux langages de programmation orientés vers la langue naturelle, le développement de l'intelligence artificielle, la reconnaissance vocale etc. Richard Parent pose la question de façon réaliste quand il écrit : «Va-t-on ériger les nouvelles structures de représentation symbolique en langue anglaise à cause du logiciel de développement qu'on utilise? Les conséquences épistémologiques ne sont-elles pas trop grandes? Aliénera-t-on la créativité et la productivité conceptuelle des usagers non-anglophones? Plus on attendra, plus le retard sur la langue anglaise sera difficile à rattraper à cause de la synergie qui s'établit à partir des quelques outils de base (syntaxe, lexiques). C'est le temps d'agir alors que commencent à se définir les composantes du marché concernant les applications de l'analyse linguistique (au sens large) dans la panoplie des outils informatiques qui seront utilisés dans les bureaux, les services au public et les résidences.» (Parent 1986:5). C'est dans le même sens que plusieurs conférenciers sont intervenus lors du Colloque de Tours en 1986, dont Jean-François Dégremont, qui a été avec Bernard Cassen un des premiers à définir le concept d'industries de la langue : «Il est clair que l'évolution ainsi amorcée va dans le sens d'une automatisation plus poussée de la manipulation des écrits (correction orthographique,

génération automatique, analyse pour résumés, assistance à la traduction) et des sons (synthèse de la parole incluant la prosodie, et reconnaissance de la parole continue multilocuteurs). Cette évolution passe nécessairement par la prise en compte des connaissances accumulées par les écrits.» (Dégremont 1986:23).

C'est dans le couple linguistique-informatique que se situent les industries de la langue génératrices des développements les plus marquants de l'informatique en 1988. Voici ce que dit Jean-François Dégremont à ce sujet : «Ainsi, linguistique et informatique ont été conduites à se rejoindre en une «informatique linguistique», également appelée «automatique linguistique» ou encore «linguistique computationnelle», visant à comprendre et représenter le fonctionnement du parlé et de l'écrit humain, afin d'en permettre l'interprétation et la génération par des automates ... Les industries de la langue sont celles qui fabriquent l'ensemble des produits dérivés des résultats de la recherche en informatique linguistique.» (Dégremont 1986:25). Le linguiste André Abbou, directeur-adjoint du Réseau des industries de la langue, en donne une définition plus synthétique dans son ouvrage récent sur les industries de la langue (Les applications industrielles du traitement de la langue par les machines) : «Nous désignerons ici par industries de la langue tous les produits, techniques, activités ou services qui appellent un traitement automatique de la langue naturelle.» (Abbou 1987:30).

C'est donc avant tout dans le domaine des industries de la langue que doit porter l'effort du français et c'est cette industrialisation qui lui permettra de demeurer encore une grande langue véhiculaire en l'an 2000. André Abbou écrit à ce sujet : «La communauté francophone, si elle veut maintenir à la langue française un usage dans les outils technolinguistiques, devra prendre en charge les coûts de traitement et d'insertion de sa langue dans de tels outils.» (Abbou 1987:22). Nous verrons quel rôle l'État devrait jouer en ce domaine de l'industrialisation du français.

A. Abbou identifie dix pôles à ce qu'il appelle le «complexe des industries de la langue» (Abbou 1987:50) :

1- la communication parlée (reconnaissance et synthèse de la parole);
2- la traduction assistée par ordinateur ou la traduction de type automatique;
3- les interfaces en langage naturel;
4- le traitement de l'écriture imprimée et manuscrite;
5- l'assistance à l'analyse, à la compréhension et à la génération des textes;
6- le traitement de texte avancé;
7- les nouveaux médias pour l'édition et l'édition de logiciels;
8- l'informatique documentaire;
9- l'édition électronique et la publication assistée par ordinateur;
10- et en surplomb à ces pôles, les dominant tous et en illustrant certains, le domaine de l'I.A. (Intelligence artificielle) s'impose comme une dimension majeure du complexe en question.

De façon concrète la maîtrise en langue nationale de ces dix aspects constitutifs des industries de la langue équivaut pour un État à dominer l'informatique de demain et participer au progrès et au modernisme planétaire. L'atteinte d'un tel objectif passe nécessairement par l'organisation de la recherche en informatique langagière selon quatre axes principaux (Abbou 1987:39) :

a) la recherche relative aux technologies de l'information et à l'informatique, destinées à l'analyse automatique du langage naturel dans sa dimension logicielle et matérielle;
b) la recherche en traitement du signal et en électronique pour la communication parlée;
c) la recherche en sciences du langage pour le traitement de l'écrit et le traitement de l'oral (génie linguistique);
d) la recherche autour des «sciences cognitives» et de l'intelligence artificielle pour l'analyse et la synthèse des mécanismes de connaisance, et notamment la compréhension du langage naturel.

Dès 1985, dans le cadre des travaux préparatoires du premier Sommet francophone (février de la même année), certains esprits précurseurs, dont Bernard Cassen, précisaient certains niveaux d'actions jugés prioritaires (*Actes de la Conférence* 1986:87) : «Développer, au sein de la Communauté francophone, les industries de la langue, c'est agir principalement à trois niveaux :

1- la formation des ingénieurs et chercheurs capables de développer et d'industrialiser les automates de manipulation du langage;
2- la recherche appliquée et l'action des pouvoirs publics ... C'est à ce niveau que sont produits les souches sur lesquelles reposent les développements industriels, banques de mots, de règles de grammaire, de sons;
3- le développement pré-industriel de produits intégrant des capacités linguistiques dans tous les secteurs d'activités.».

Le facteur linguistique est mis au premier plan dans la réalisation de cette entreprise, comme on peut le lire dans les Actes du Sommet de Paris: «Ainsi donc, le facteur linguistique, en lui-même mineur, est érigé en substrat et en vecteur de la production et de l'usage des données de toutes sortes. Derrière l'apparence anodine du fait linguistique se profile donc soit la capacité d'une nation ou d'un groupe de nations à créer et à produire des biens industriels conditionnant l'ensemble de leurs activités, soit un statut de simple consommateur, dans la langue de l'autre, de produits stratégiques.» (*Actes de la Conférence* 1986:86).

C'est à la suite des réflexions continues menées depuis 1980 et pilotées par certains États francophones, comme la France et le Québec, déjà bien impliqués dans la recherche en «génie linguistique» et aux représentations des groupes de recherches oeuvrant en ce domaine, que la structure des Sommets francophones née en 1986 constituait un Réseau des industries de la langue pour la promotion des intérêts de la langue française en ce domaine et affirmait par là la volonté des Chefs d'État francophones d'en faire une priorité. Animé par la France, ce Réseau réunit régulièrement depuis 1986 des responsables et des experts nationaux francophones de plusieurs États représentés au Sommet pour préparer des propositions à l'intention des Chefs d'États réunis lors des Sommets. L'antenne québécoise de ce Réseau, constituée par le Sous-comité québécois des industries de la langue a joué et joue toujours le rôle de rassembleur et d'animateur pour le Sommet de projets susceptibles de développer les industries de la langue à l'échelle de la francophonie. C'est ainsi qu'à la suite du premier Sommet de Paris, le Sous-comité québécois se mettait au travail pour préparer pour le Sommet de Québec (septembre 1987) des propositions plus précises touchant les industries de la langue. Une proposition contenant trois programmes majeurs a été élaborée par le Comité et avaliséé par le Réseau (*Rapport du sous-comité des IDLL* 1987). Le premier touche la recherche et le développement industriel dans les domaines de la communication parlée, du traitement de l'écrit et de l'intelligence artificielle (financement multilatéral de 15 millions (NF), financement national de certains projets), proposition assortie de la création d'un Observatoire des évolutions technologiques, des innovations et des marchés liés aux industries de la langue et d'un Forum des industries de la langue. Le second programme est celui de la Création d'un Réseau international de néologie et de terminologie pour faciliter l'adaptation de la langue française au modernisme technologique. Le troisième, enfin, est orienté vers la formation et le perfectionnement de chercheurs et de praticiens dans le domaine des industries de la langue. Ce troisième programme met l'accent sur le développement d'un axe Nord-Sud s'inscrivant dans une démarche globale «industrialisante» pour que «la langue française devienne également une langue relais servant au transfert des connaissances scientifiques et techniques et que, de cette façon, les autres langues de l'aire francophone atteignent progressivement un niveau de performance propre à satisfaire les besoins des pays du Sud.» (*Rapport du sous-comité des IDLL* 1987). Ces propositions adoptées durant le Sommet de Québec devraient être appuyées financièrement par un fonds multilatéral de quelque 25 millions (NF) et l'engagement des chefs d'État à favoriser la recherche en ce domaine dans leurs États respectifs.

Dans les faits, on attend encore depuis le premier Sommet de Paris et celui de Québec en 1987 des réalisations concrètes en ce domaine. D'une part les financements prévus sont constamment revus à la baisse par voie diplomatique (l'un attend la mise de l'autre) et rarement dégagés dans ce cadre.

D'autre part, l'organisation des Sommets (Comité du suivi) est davantage préoccupée par la préparation du prochain Sommet (Dakar en ce moment) en favorisant des petits projets «visibles» et peu coûteux qui alimenteront cet évènement, ignorant par là la nécessité d'asseoir les programmes pris en charge par les Sommets sur des bases permanentes moins fluctuantes selon les politiques du moment. On peut toutefois espérer que dans les prochaines années, des réalisations notoires à ce titre pourront être conduites en aplanissant les obstacles de concurrence économique entre les États (les industries de la langue sont génératrices de revenus et développeront encore davantage dans les prochaines années cet aspect - il ne faut pas l'oublier).

Une chose est certaine, une organisation panfrancophone de la recherche dans le domaine des industries de la langue est devenue nécessaire pour assurer l'avenir du français; une organisation nationale est un prérequis à une telle démarche et est donc tout aussi essentielle. Il convient de souligner la nécessité du développement développement d'une terminologie française de l'informatique qui soit fonctionnelle et qui favorise par le fait même le dialogue entre experts et les échanges internationaux en langue française. Enfin, il faudra aplanir un autre obstacle matérialisé par la variation linguistique du français et faire en sorte que les machines puissent maîtriser cette variation en respectant les variétés de français telles qu'elles se pratiquent dans les diverses régions de la francophonie. A ces conditions, le français a toutes les chances de demeurer une grande langue véhiculaire capable de transmettre les savoirs modernes et d'assurer à ses usagers la possiblité de tirer profit du fait d'être et de demeurer francophone.

BIBLIOGRAPHIE

ABBOU, André, **MEYER**, Thierry, **LEFAUCHER**, Isabelle. *Les industries de la langue. Les applications industrielles du traitement de la langue par les machines*, Paris, Éditions Daicadif, 1987, 400 p.

Actes de la Conférence des chefs d'État et de Gouvernement des pays ayant en commun l'usage du français, La Documentation française, 1986.

BARANES, William. **Les industries de la langue**, *Qui vive international*, n° 4, pp. 74-75.

CALVET, Louis-Jean. 1981. *Les langues véhiculaires*, coll. Que sais-je, no 1916, PUF, Paris, 128p.

CASSEN, Bernard. 1986. **Un nouveau front pour le français et la langue de l'Europe**, LES INDUSTRIES DE LA LANGUE. ENJEUX POUR L'EUROPE. Actes du colloque de Tours, Tours, 28 fév. - 1er mars 1986, dans *Encrages*, n° 16, 12-14.

DEGREMONT, Jean-François. **Au croisement de l'informatique et de la linguistique**, LES INDUSTRIES DE LA LANGUE. ENJEUX POUR L'EUROPE. Actes du colloque de Tours, Tours, 28 fév. - 1er mars 1986, dans *Encrages*, n° 16, nov. 1986, pp. 22-46.

Documents de conférence. 1987. Deuxième conférence des chefs d'État et de gouvernement ayant en commun l'usage du français (2-4 septembre 1987), Québec, Canada, 191 p.

DALBY, David, **CORBEIL**, Jean-Claude, **ARNOLD**, Thierry. 1985. *Les langues et l'espace du français*, ACCT-CIRELFA, Publications CIRELFA, 54p.

HAGEGE, Claude. 1987. *Le français et les siècles*, Editions Odile Jacob, Paris, 271 p.

La francisation en marche. 1986. «L'informatique en français, un atout majeur pour la conquête du marché mondial», vol. 5, no 6, 15 p.

LANG, Jacques. 1986. **Contre un espéranto au rabais: les ordinateurs au service de la diversité**, LES INDUSTRIES DE LA LANGUE. ENJEUX POUR L'EUROPE, Actes du colloque de Tours, Tours, 28 fév. - 1er mars 1986, dans *Encrages*, n° 16, 17-20.

OUELLON, Conrad. 1986. *L'informatique en français*, texte d'une conférence présentée à Québec le 21 mars 1988 devant la Société Saint-Jean Baptiste, 17 p. dactylographiées.

PARENT, Richard. *Le français en informatique à l'heure des systèmes experts*, texte d'une allocution présentée devant les membres de l'Association de la recherche en communication du Québec (Montréal 8 et 9 mai 1986), Québec, 1986, 15 p. dactylographiées.

Rapport d'étape du Comité Canada-Québec sur le développement du logiciel d'expression française: sommaire et faits saillants, Québec-Canada, juin 1986, 6 p. dactylographiées.

Rapport de synthèse: Industries de la langue. Conférence des chefs d'État et de Gouvernement des pays ayant en commun l'usage du français, Paris, 17 - 19 fév. 1986, *Actes de la Conférence des chefs d'État et de Gouvernement des pays ayant en commun l'usage du français*, La Documentation française, 1986, pp. 85-90.

Rapport du Sous-comité des Industries de la langue au Comité de coordination des contributions québécoises, Secrétariat québécois au Sommet de Québec, 1987, 100 p.

L'ÉVALUATION DES DIDACTICIELS: ÉTAT ET NATURE DE LA DOCUMENTATION[1]

France H.-Lemonnier et Marie-Josée Fortin[2]
Langues et linguistique
Université Laval

Ce travail s'inscrit dans le cadre d'un projet intitulé «Évaluation, analyse et élaboration d'un didacticiel de langue»[3] lui-même faisant partie d'un projet plus vaste, le projet RELAI[4]. Il constitue en quelque sorte un résultat préliminaire du premier volet de notre recherche[5].

Le domaine de l'évaluation, appliqué à l'enseignement / apprentissage par ordinateur (EAO), est plutôt jeune puisque la documentation présentée dans le cadre de cette étude a été publiée entre 1975 et 1988.

Par ailleurs, il peut paraître inhabituel de s'intéresser à un tel sujet, à savoir la présentation de l'état et de la nature d'une documentation donnée. Mais si l'on tient compte, d'une part, de la nouveauté du domaine des EAO autant pour les didacticiens que pour les linguistes et, d'autre part, du nombre restreint et coûteux des didacticiels actuellement disponibles sur le marché, ce travail pourrait en intéresser plus d'un, notamment ceux qui oeuvrent dans le domaine des EAO à titre d'utilisateurs ou de futurs concepteurs, tout en ayant une connaissance restreinte du domaine de l'évaluation.

L'objectif de cette étude sera donc de proposer une classification générale de ces ouvrages et d'exposer quelques facettes de leur contenu. Nous présenterons successivement les caractéristiques des 35 instruments d'évaluation répertoriés (I), celles des 93 articles ou monographies (II), et de façon très brève, celles d'une centaine de catalogues de promotion de didacticiels (III).

Nous tenterons de faire ressortir les traits dominants et distinctifs de ces trois répertoires, en apportant quelques points de vue critiques.

---

[1]Communication présentée lors du 56e Congrès de l'ACFAS, Moncton, 9-13 mai 1988.

[2]Nous tenons à remercier Madame Renate Moisan pour sa précieuse collaboration à toutes les étapes de ce projet : constitution des répertoires, analyse des documents, hypothèses de classification et révision du contenu de cette communication.

[3]Le projet «Évaluation, analyse et élaboration de didacticiels» comporte trois volets. Le premier vise à faire une synthèse des ouvrages portant sur l'évaluation des didacticiels. Le second a trait à la description détaillée des caractéristiques du didacticiel «idéal» pour l'enseignement et l'apprentissage du français langue seconde ou étrangère. Le troisième sera consacré à l'élaboration d'un prototype immédiatement suivi de son expérimentation. Le premier volet a pu être réalisé grâce à une subvention accordée par le projet RELAI.

[4]Le projet RELAI (Recherche en linguistique appliquée à l'informatique) s'inscrit dans le cadre d'une entente entre l'Université Laval et la Société IBM.

[5]Voir «Les données de l'évaluation des didacticiels : essai de synthèse et prospectives pour l'EAO-langues», rapport de rechercherédigé par les co-responsables du projet, Mmes France H.-Lemonnier et Renate Moisan.

I. Caractéristiques de 35 instruments d'évaluation des EAO[6]

Nous avons choisi de présenter les caractéristiques de ces instruments selon deux points de vue. Le premier, classification générale des instruments (1.1.), se consacre exclusivement à la description externe de ces documents, à savoir les objectifs d'évaluation, la clientèle visée par ces instruments, le contexte de travail de leurs auteurs et enfin s'ils sont conçus pour évaluer une matière spécifique.

Le second, contenu général des instruments (1.2.), permet plutôt de faire une description interne et de dégager le type de questionnement proposé, en fonction cependant d'un seul paramètre, l'orientation des auteurs de ces instruments.

1.1. *Classification générale des instruments*

La classification générale de ces 35 instruments repose essentiellement sur quatre critères: l'objectif d'évaluation (1.1.1.), la clientèle visée (1.1.2.), le contexte de travail de leurs auteurs (1.1.3.) et la matière spécifique à évaluer (1.1.4.).

1.1.1. L'objectif d'évaluation

Ces 35 instruments peuvent d'abord se distinguer par l'objectif d'évaluation, que Owston (1986) qualifie de formelle ou d'informelle. Ainsi, l'utilisation d'un instrument d'évaluation formelle se fait selon une procédure rigoureuse qui conduit à la recommandation ou à la non-recommandation du didacticiel examiné.

Par ailleurs, un instrument d'évaluation informelle ne peut être utilisé qu'à des fins personnelles et sa procédure d'utilisation est moins structurée.

Parmi les 35 instruments examinés, six seulement conduisent à une évaluation formelle.

1.1.2. La clientèle visée

Ces instruments s'adressent principalement et parfois simultanément à trois clientèles. Ils sont d'abord destinés à l'évaluateur, c'est-à-dire celui qui est chargé d'observer ou d'évaluer un produit. Ils peuvent également être destinés à des concepteurs et à des utilisateurs, enseignants ou non.

Cependant, les auteurs de ces différents instruments ont souvent donné une interprétation plus large au terme «évaluateur». Pour certains, l'évaluateur peut être un directeur d'institution, un enseignant ou toute personne responsable du dossier «sélection et achat de didacticiels» ou un évaluateur au sens strict. C'est pourquoi un instrument a pu être conçu avec plus d'un objectif, dont celui d'aider le concepteur à vérifier certaines étapes de l'élaboration d'un produit, tout en permettant à des enseignants, producteurs, distributeurs, utilisateurs, d'évaluer un didacticiel.

1.1.3. Le contexte de travail des auteurs

Les auteurs de ces instruments sont essentiellement de quatre types. Douze d'entre eux sont des universitaires, c'est-à-dire des professeurs-chercheurs attachés à différents établissements universi-

[6]Les références de ces 35 instruments n'apparaissent pas dans le présent texte. Elles seront par ailleurs complètes dans le rapport de recherche relatif au premier volet de ce projet (voir note précédente).

taires. Douze, également, travaillent pour un organisme gouvernemental, soit dans des ministères de l'Éducation de différentes provinces du Canada ou de différents États américains. Huit oeuvrent plutôt à titre de chercheurs affiliés à différents organismes, tels le **National Council of Teachers of Mathematics**, l'**Institut d'études pédagogiques** de l'Ontario, **MicroSIFT**, **Epie**, **NEA**, etc. Les trois derniers proviennent d'organismes commerciaux.

1.1.4. La matière spécifique à évaluer

Un seul instrument a été spécifiquement conçu pour évaluer des didacticiels de mathématiques. Cinq visent l'évaluation de didacticiels de langues, langue maternelle ou langue seconde ou étrangère. Pour les 29 autres, les auteurs n'ont pas mentionné si l'instrument était conçu pour évaluer un didacticiel visant l'enseignement d'une matière donnée.

Cette classification générale permet de constater que les instruments d'évaluation des EAO peuvent être utilisés à plus d'une fin et par des clientèles différentes, quelle que soit la matière enseignée. Toutefois, un nombre relativement restreint d'instruments permet de porter réellement un jugement sur les didacticiels examinés.

1.2. *Contenu général des instruments*

L'examen plus détaillé du contenu de ces instruments permet de dégager que l'observation et l'évaluation d'un didacticiel donné doit se faire en deux moments, l'un pouvant avoir lieu sans visionner le produit, l'autre l'exigeant.

Il est par ailleurs possible d'identifier trois types de questionnement. Le premier s'adresse plutôt à la documentation du produit à examiner (1.2.1.), le second à la performance générale du produit (1.2.2.) et le troisième, à la performance du produit en regard des utilisateurs (1.2.3.).

Chaque questionnement peut conduire à un jugement plus ou moins élaboré.

1.2.1. Le questionnement de la documentation

Le questionnement de la documentation permet d'obtenir trois types d'informations sur le didacticiel examiné, soit des *informations très générales*, telles le titre, l'auteur, l'année de publication, le prix, le producteur, la matière enseignée, le type de didacticiel et la clientèle visée, soit *des informations d'ordre technique*, notamment le modèle et la marque du micro-ordinateur sur lequel fonctionne le didacticiel, la mémoire requise, la disponibilité de copies de secours, le type de fonctionnement à savoir sur disque ou sur réseau, le langage de programmation utilisé, le support à l'implantation s'il s'agit d'un produit plus complexe, le type de matériel et de logiciel requis pour son utilition maximale. Il permet enfin d'obtenir *des informations plus spécifiques* sur le contenu, sur le modèle pédagogique retenu et sur les objectifs du didacticiel.

En somme, le questionnement de la documentation permet à l'évaluateur de juger si la description générale du didacticiel est plus ou moins conforme à l'attente d'un consommateur éventuel.

1.2.2. Le questionnement de la performance du didacticiel

Le questionnement de la performance cherche à savoir si le produit fonctionne bien, s'il contient ce qu'on a préalablement décrit. Pour évaluer cet aspect, l'évaluateur doit visionner le didacticiel et tenter de répondre à des questions portant sur le contenu, les objectifs, l'intérêt pédagogique, la pertinence d'utiliser l'ordinateur pour enseigner cette matière.

Les questions sont conçues pour permettre de juger si le contenu présenté est exact, s'il convient au niveau suggéré, s'il est conforme au programme scolaire, s'il est exempt d'erreurs et si les objectifs énoncés sont atteints.

Par ailleurs, certaines questions permettent de mesurer la performance informatique proprement dite, à savoir la facilité du branchement, la fiabilité, la souplesse, l'efficacité des structures d'aide, la rétroaction, de même que ses qualités techniques (couleurs, son, graphisme, déroulement des affichages, des pages-écran, etc.).

Ce type de questionnement est sans aucun doute le plus exigeant. Il recouvre les aspects les plus importants de l'évaluation d'un enseignement/apprentissage par ordinateur, soit le contenu, les objectifs, le niveau, l'intérêt pédagogique et la performance informatique proprement dite. Il rejoint en quelque sorte les préoccupations premières de tout enseignant.

1.2.3. Le questionnement de la performance du produit en regard des utilisateurs

Le questionnement de la performance du produit en regard des utilisateurs contient un certain nombre de questions permettant de juger si le didacticiel est orienté vers l'utilisateur, selon les points de vue de l'enseignant et de l'apprenant.

On cherche entre autres à savoir si l'enseignant doit intervenir auprès de l'apprenant, si l'intérêt des apprenants est maintenu et si ces derniers contrôlent facilement le déroulement du logiciel, et enfin si l'âge et le niveau suggérés par les producteurs convient bien aux utilisateurs observés.

Autrement dit, ce type de questionnement permet de mesurer si l'intérêt de l'apprenant est soutenu, si le contenu convient bien à l'âge et au niveau suggérés et si l'utilisateur a une certaine facilité à contrôler le déroulement technique du didacticiel.

D'autre part, aucun de ces questionnements ne permet de mesurer l'efficacité d'un apprentissage fait au moyen d'un didacticiel donné en regard d'un autre moyen. Il ne permet pas non plus de mesurer l'efficacité de l'utilisation de l'ordinateur pour l'enseignement/apprentissage d'une matière donnée, en l'occurrence l'enseignement/apprentissage des langues.

Il permet tout au plus de se faire une idée générale des qualités minimales de base que devrait posséder un didacticiel.

II. Caractéristiques de 93 articles et monographies portant sur l'évaluation des EAO

Dans l'ensemble de la littérature portant sur l'évaluation de didacticiels, une place importante est également accordée aux instruments d'évaluation. En effet, ceux-ci sont souvent analysés ou critiqués, toujours dans le but d'améliorer le processus d'évaluation, pour ainsi assurer une meilleure investigation de la qualité des didacticiels. Les ouvrages s'intéressant aux instruments d'évaluation ne constituent cependant qu'une des cinq catégories selon lesquelles il est possible de les classifier.

Nous présenterons cette partie de notre documentation selon les points de vue adoptés lors de la présentation des instruments (I). Le premier, classification générale (2.1.), se consacre à leur description externe, à savoir leur type, leur provenance et le contexte de travail de leurs auteurs. Le second, contenu général (2.2.), s'attache à leur description interne et permet de dégager quelques facettes de ces écrits.

2.1 *Classification générale des articles et monographies*

Les ouvrages sont en majorité tirés de périodiques à caractère éducatif, issus pour la plupart de différentes associations. Ils proviennent des États-Unis, du Québec et de l'Ontario; il y en a très peu de l'Europe.

Les auteurs sont généralement des chercheurs affiliés à des universités, ou rattachés à des organismes gouvernementaux et commerciaux. Notons enfin que parmi l'ensemble de ces ouvrages, seulement six s'intéressent aux langues de façon spécifique.

En somme, comme pour d'autres domaines, les monographies se font plus rares que les articles, et comme on pouvait s'y attendre, les États-Unis sont en tête de liste en ce qui a trait au nombre d'écrits. Il est également intéressant de constater que le contexte de travail des auteurs de ces écrits est le même, qu'ils s'agissent d'articles, de monographies ou d'instruments. Enfin, le domaine de l'enseignement des langues proprement dit n'a guère occupé de place d'un point de vue quantitatif, puisqu'on n'a dénombré que douze écrits, soit six articles et six instruments.

2.2 *Contenu général des articles, documents et communications*

Le contenu général des articles, documents et communications s'est avéré plus difficile à circonscrire que celui des instruments, compte tenu de la prise en compte d'éléments aussi variés que des modèles d'évaluation, des modèles de didacticiels, des stratégies d'apprentissage, des critères de qualité, des procédures à adopter pour cerner les critères d'un bon didacticiel, des projets d'évaluation, des théories d'enseignement/apprentissage dont il serait souhaitable de tenir compte, etc. Malgré la diversité de ces points de vue et pour les besoins de cette étude, nous présentons, sous toutes réserves, une première classification en cinq catégories. La première concerne les positions prises en regard de la qualité des didacticiels (2.2.1.), la seconde, les approches d'évaluation formative (2.2.2.), la troisième, les positions prises en regard des instruments d'évaluation (2.2.3.), la quatrième, les opinions et les commentaires généraux (2.2.4.) et la cinquième, l'historique de l'évaluation (2.2.5.).

2.2.1. La qualité des didacticiels

Une quarantaine d'articles traitent des critères de qualité inhérente ou souhaitée à tout bon didacticiel, selon différentes perspectives dont les unes s'attachent à des aspects plus spécifiques.

Ainsi, certains auteurs (Cruickshand et Telfer 1980 et Bell 1985) s'intéressent à l'évaluation de didacticiels en fonction du type d'interaction (jeux et simulation, tutoriel, exerciseur, etc.) ou du type d'exercices (exercices de discrimination, application de règles, solution de problèmes, etc.).

D'autres (McCan 1981 et Rashio et Lange 1984) prônent l'utilisation de stratégies d'apprentissage telles le contrôle du programme par l'apprenant l'amenant à de fréquentes prises de décisions, ou encore, pour ceux qui ont besoin d'une aide extérieure, le travail à deux ou trois sur le même terminal.

D'autres encore mentionnent la trop grande importance accordée par certains auteurs aux caractéristiques techniques ou à la qualité du didacticiel et proposent des critères plus variés (Budoff, Thormann et Gras 1985). Muller (1985), pour sa part, présente un ensemble de critères pouvant être utilisés à titre expérimental dans des recherches en évaluation de didacticiels.

D'autres suggèrent des stratégies permettant de résoudre les désaccords relatifs aux critères d'un bon didacticiel (Roblyer 1981), ou des méthodes systématiques pour évaluer des didacticiels (Sakamoto et autres, 1979) ou des procédures permettant d'extraire des critères utiles (Jack 1983).

Il existe par ailleurs des perspectives plus globales, et plus englobantes, comme l'approche de Della-Piana (1982), mettant en parallèle la critique de film et l'évaluation de didacticiels ou celle de Roblyer (1981) qui introduit le concept de «standards», donnant lieu à un ensemble de critères répartis en trois catégories: caractéristiques essentielles, caractéristiques esthétiques et caractéristiques propres à la matière enseignée. Blum-Cohen (1983) et Nesbit (1984) vont dans le même sens et soulignent le besoin de déterminer ce qu'est un bon didacticiel. D'autres auteurs (Wade 1980; Wager 1981 et Steffin 1983) poussent plus loin le besoin d'établir des caractéristiques propres à la matière enseignée en prônant la prise en compte d'une «théorie de base», prenant appui sur les principes d'apprentissage de Gagné et Briggs (1979). Compte tenu que chaque matière enseignée requière des conditions d'apprentissage différentes, une analyse préalable des apprenants et des concepts à enseigner permettra en effet de déterminer les principes essentiels à un apprentissage efficace.

2.2.2. Les approches d'évaluation

Sous cette catégorie, figurent vingt-deux articles traitant de l'évaluation formative et/ou sommative, distinction établie par Scriven (1967). L'évaluation formative est effectuée durant le développement d'un produit et avant sa mise en marché, et l'évaluation sommative, lorsque le produit est complété.

La plupart des auteurs ayant écrit sur le sujet (Truett 1984; Golas 1983 et Avner 1975) relatent la difficulté d'utiliser une procédure d'évaluation formative en tenant compte de ses véritables fins, quel que soit le matériel didactique à évaluer. Seuls Brown et Ellinger (1978) présentent une procédure d'évaluation formative et sommative de didacticiels. L'étude de McMillan (1981) traite de l'évaluation de matériel écrit.

2.2.3. Les instruments

Les articles regroupés sous cette catégorie constituent une présentation sous forme d'analyse ou de revue critique de certains instruments d'évaluation. Certains ont fait des études comparatives ou des synthèses (St-Laurent 1984; Ashmore 1984 et Wager 1981, etc.)

D'autres (Miller et Burnett 1984; Jonhston 1987, etc.) font des mises en garde relatives aux contextes de leur utilisation. Enfin, Rashio et Lange (1984) soulignent le besoin d'élaborer des instruments d'évaluation pour chaque discipline, en particulier pour les didacticiels de langues.

2.2.4. Les opinions et commentaires

Sans fondement particulier et sans argumentation très élaborée, certains auteurs, dans de courts articles, présentent des considérations d'ordre général, souvent subjectives. Ils donnent parfois des conseils judicieux ou font des prédictions qui amènent à la réflexion.

2.2.5. L'historique de l'évaluation

Les principales méthodologies ou approches pour l'évaluation de didacticiels sont bien dégagées, et selon l'ordre chronologique, dans les articles de Borich et Jemelka (1981) et Johnston (1987). Dans la même perspective, Dudley-Marling et Owston (1987) s'intéressent à l'évaluation en regard de différents critères sur laquelle elle est basée.

Cette présentation très sommaire du contenu général de cette partie de la documentation des EAO permet de constater la grande variété des thèmes abordés et des perspectives envisagées. Toutefois, le besoin de définir ce qu'est un bon didacticiel et le besoin de s'appuyer sur une théorie de base ressortent nettement, que ce soit pour élaborer un didacticiel ou une procédure d'évaluation de didacticiels.

III. Caractéristiques d'une centaine de catalogues de promotion de didacticiels

Des 273 lettres envoyées dans le but de nous procurer des catalogues de promotion de didacticiels, 131 sont demeurées sans réponse, ce qui s'explique par les nombreux changements d'adresse des maisons d'édition et aussi par la disparition de plusieurs d'entre elles.

Nonobstant ces faits, nous avons réussi à obtenir 194 catalogues provenant principalement des États-Unis et du Canada, quelque-uns de la France. Cependant, nous avons dénombré seulement 84 catalogues qui faisaient la promotion de 99 didacticiels de langues.

L'un des objectifs de cette recherche était de réaliser une liste de didacticiels de français langue seconde pour étudiants non-francophones en milieu universitaire.

Par ailleurs, il a été très difficile d'arrêter notre choix sur ces 99 didacticiels, puisque les descriptions qui en sont faites dans les catalogues ont un caractère plus commercial que pédagogique. Elles s'adressent au grand public, et sont souvent peu ou pas représentatives, aussi bien du contenu que de la démarche pédagogique.

Les résultats présentés dans le cadre de cette étude ne concernent toutefois que le public et le niveau, puisque ceux-ci sont les premiers critères à considérer lors de l'achat de tels produits. Or, selon ces deux critères, il s'avère que, parmi ces 99 didacticiels, 34 sont conçus pour des étudiants, soit de niveaux collégial (6), débutants, intermédiaires, avancés (10), de la 1re secondaire au collégial (6), de la 1re secondaire à l'université (7), de la 4e secondaire à l'université (1) ou encore pour des étudiants de tous niveaux (4).

Pour 14 d'entre eux, aucun niveau n'est mentionné; ils peuvent ainsi s'adresser à des adolescents, à des adultes ou à tous les âges. Les 51 suivants ne s'adressent pas à un public spécifique et devraient convenir à tous les niveaux, débutants, intermédiaires ou avancés.

Ainsi, sur un total de 99 didacticiels, 30 sont destinés à une certaine clientèle de niveaux variés.

Ces résultats amènent à constater qu'il est difficile, voire même impossible, de sélectionner des didacticiels à partir de telles descriptions. L'effet de promotion des catalogues est alors considérablement diminué. Leur seul avantage réside dans le fait de nous faire connaître l'existence de titres et de maisons de distribution. Pour connaître un peu mieux ce type de produits, il faut alors s'en procurer, sans savoir s'ils répondent à nos besoins, vu l'impossibilité de les visionner.

Le manque de transparence des descriptions nuit sans aucun doute à la promotion de ce type de produits et par conséquent, aux concepteurs-producteurs. De plus, le fait que plusieurs didacticiels

puissent convenir à plusieurs clientèles, du secondaire à l'université, à titre de débutants en langue, intermédiaires ou avancés, laisse un peu perplexe si l'on prend en compte notamment le seul critère de la motivation.

Synthèse et conclusion

Les instruments d'évaluation proposent des questionnements suffisamment élaborés pour fournir des critères de choix à un utilisateur, ou pour informer un futur concepteur des dimensions dont il devrait tenir compte, lors de l'élaboration d'un didacticiel. Cependant, les questions recouvrent des champs tellement généraux qu'il est possible de faire l'évaluation de didacticiels de mathématiques et de didacticiels de langues avec un même instrument.

Les articles, documents et communications portant sur l'évaluation des didacticiels permettent de mieux circonscrire la variété des perspectives prises en compte par ce domaine. Une analyse sommaire du contenu de cette documentation permet en effet de dégager certaines priorités, soit à l'égard de la classification des nombreux critères proposés dans les instruments, soit à l'égard du besoin de s'appuyer sur des théories lors de la conception de modèles.

Les catalogues nous font connaître quantité de titres, de maisons d'édition et de prix.

Cette documentation pourra en outre être utile aux enseignants. D'une part, elle leur permettra de se familiariser avec les concepts véhiculés dans l'évaluation de didacticiels. D'autre part, elle leur permettra de développer une certaine vigilance lors de visionnements de didacticiels.

De plus, d'un certain point de vue, cette documentation est trop élaborée pour ce que sont par exemple certains didacticiels examinés, la plupart étant des programmes courts et recouvrant seulement un aspect d'une matière donnée. D'autre part, elle est trop restreinte puisqu'elle permet difficilement de mesurer des aspects importants, notamment la pertinence et l'efficacité de l'utilisation de l'ordinateur pour l'enseignement/apprentissage d'une matière donnée.

Par ailleurs, il est possible de dégager de cette documentation certains aspects «théoriques» caractérisant un didacticiel de qualité, - en dépit du fait que l'enseignement/apprentissage varie en fonction du type d'étudiants et de leurs besoins, de la sorte de contenu et des objectifs spécifiques du matériel.

Ces aspects seraient les suivants :

1- Un bon didacticiel repose sur l'élaboration d'un «design» général et de différents «designs» spécifiques tels celui du contenu à enseigner, celui des activités d'apprentissage, celui des tests, celui de la présentation, celui de la performance technique, etc.
2- Lors de l'élaboration de tels «designs», l'équipe de concepteurs s'est appuyée sur des théories, quel que soit le design à élaborer.
3- Un bon didacitciel a été soumis à une procédure d'évaluation formative (pendant son développement) et sommative (avant sa mise en marché définitive).

Cependant, avant d'entreprendre l'élaboration d'un didacticiel de français langue seconde, destiné à des étudiants non-francophones en milieu universitaire (ce qui constitue l'un des volets de notre projet), des réflexions préalables s'imposent, relativement à la pertinence et à la rentabilité d'un tel projet.

Supposons que l'analyse de la pertinence s'avère positive, on peut en outre se demander si le coût d'investissement en vaudrait la peine compte tenu qu'un tel enseignement nécessite l'achat de plusieurs postes d'ordinateur. Pourrait-on rentabiliser un tel investissement au point de vue économique, étant donné que les universités ne sont pas régies par des programmes uniformes?

Il nous semble qu'il faudrait pouvoir répondre à de telles questions, et à bien d'autres, avant de se lancer dans un projet d'élaboration de didacticiels. La prise en compte des éléments de la documentation présentée dans le cadre de cette étude constituerait sans doute un sage point de départ.

BIBLIOGRAPHIE

ASHMORE, T.M. (1984). "Evaluating CAI Material for the Microcomputer". Paper presented at the Annual Meeting of the Speech Communication Association, Chicago, [ERIC ED-252180].

AVNER, R.A. (1975). "The Evolutionary Development of CAI Evaluation Approaches". Paper presented at the Annual Meeting of the American Educational Research Association, Washington, D.C. [ERIC ED-105897].

BELL, M.E. (1985). "The Role of Instructional Theories in the Evaluation of Microcomputer Courseware". In *Educational Technology*, 25 (3), 36-40.

BLUM-COHEN, V.L. (1983). "Criteria for the Evaluation of Microcomputer Courseware". In *Educational Technology*, 23 (1), 9-14.

BORICH, G.D. et R.P. JEMELKA (1981). "Evaluation". In *Computer-Based Instruction: A State-of-the-Art Assessment*, edited by Harold F. O'Neil Jr., New-York: Academic Press, chap. 6, 161-209.

BROWN, B.R. et R.S. ELLINGER (1978). "Evaluation of Computer-Based Instructional Material: Intent and Procedures." In *ACM TOPICS in Instructional Computing. A Special Publication of SIGCUE*, 22-23.

BUDOFF, M., THORMANN, J. et A. GRAS (1985). "Evaluating and Seclecting Software". In *Microcomputer in Special Education*, Cambridge / Massachusetts: Brookline Books, 129-139.

CRUICKSHAND, D.R., et R. TELFER (1980). "Classroom games and simulations". In *Theory into Practice*, 19 (1), 75-80.

DELLA-PIANA, G.M. (1982). "Film Criticism and Micro-Computer Courseware Evaluation". In *New Directions for Program Evaluation: Field Assesments of Innovative Evaluation Methods*, edited by N.L. Smith, San Francisco: Jossey-Bass Publishers, 11-28.

DUDLEY-MARLING, C. et R.D. OWSTON (1987). "The State of Educational Software: A Criterion-Based Evaluation". In *Educational Technology*, 27 (3), 25-29.

GAGNE, R.M. et L.J. BRIGGS (1979). In *Principles of Instructional Design*, New York: Holt, Rinehart, and Winston.

GOLAS, K.C. (1983). "The Formative Evaluation of Computer-Assisted Instruction". In *Educational Technology*, 23 (1), 26-28.

JACK, C.S. (1983). "Effective Evaluation of Software for Resale". In *The DES Professional*, Professional Press Inc., 2 (6), 112, 114, 170.

JONHSTON, V.M. (1987). "The Evaluation of Microcomputer Programs: an Area of Debate". In *Journal of Computer-Assisted Learning*, 3 (1), 40-50.

MCCAN, P.H. (1981). "Learning Strategies and Computer-Bnased Instruction". In *Computers and Education*, 5 (2), 133-140.

MCMILLAN, R.A. (1981). "Formative Evaluation of Textbooks". In *Evaluation News*, 2 (4), 343-349.

MILLER, L. et J.D. BURNETT (1986). "Theoretical Considerations in Selecting Language Arts Software." In *Computers and Education*, 10 (1), 159-165.

MULLER, E.W. (1985). "Application of Experimental and Quasi-experimental research Designs to Educational Software Evaluation". In *Educational Technology*, 25 (10), 27-31.

NESBIT, I.S. (1984). "Evaluating Micro Software". In *Datamation*, 30 (11), 74-78.

OWSTON, R.D. (1986). *Software Evaluation. A Criterion-Based Approach*, Toronto: York University Press, 81 p.

RASHIO, R. et D.L. LANGE (1984). "A Discussion of the Attributes, Role, and Uses of CAI Materials in Foreign Languages". In *Journal of Computer-Based Instruction*, 11 (1), 22-27.

ROBLYER, M.D. (1981). "When is it "Good Courseware?" Problems in Developing Standards for Micro-computer Courseware". In *Educational Technology*, 21 (10), 47-54.

SAINT-LAURENT, L. (1984). *L'évaluation de didacticiels: document provisoire*, Ste-Foy: INRS-Education, [EDUQ-8570679].

SAKAMOTO, T., KIMURA, K., SHIMADA, M., OZAWA, S., and T. OKAMOTO (1979). "Evaluation Methods for Improving CAL Materials". In *Computers and Education*, 3 (4), 281-291.

SCRIVEN, M. (1967). "The Methodology of Evaluation". In R.W. Tyler, R.M. Gagne, and Scriven, M. (Eds). *Perspectives of Curriculum Evaluation*, Chicago: Rand McNally.

STEFFIN, S.A. (1983). "A Suggested Model for Establishing the Validity of Computer-Assisted Instructional Materials". In *Educational Technology*, 23 (1), 20-22.

TRUETT, C. (1984). "Field Testing Eductional Software: Are Publishers Making the Effort?". In *Educational Technology*, 24 (5), 7-12.

WADE, T.E. Jr. (1980). "Evaluating Computer Programs and Other Teaching Units". In *Educational Technology*, 20 (11), 32-35.

WAGER, W. (1981). "Issues in the Evaluation of Instructional Computing Programs". In *Educational Computer Magazine*, 1 (3), 20-22.

PHONO:
UN APPLICATEUR DE RÈGLES PHONOLOGIQUES CATÉGORIQUES ET VARIABLES[1]

Claude Paradis
Langues et linguistique
Université Laval

I Introduction

Ce projet a vu le jour dans le cadre d'un séminaire d'automatisation linguistique que nous avons donné à l'université Laval à l'automne 1987. Il avait pour but d'initier les étudiants qui y assistaient à certaines notions d'algorithmique et de programmation en linguistique et plus particulièrement en phonologie et en phonétique.

Le projet a ensuite pris plus d'ampleur quand nous avons décidé de tenter de modéliser une partie des connaissances que les phonologues, les phonéticiens et les sociolinguistes avaient accumulées depuis un certain nombre d'années.

II Plan de l'exposé

L'exposé portera sur les points suivants : d'abord, une explication des objectifs poursuivis dans le cadre de cette recherche et de la structure générale du programme tel qu'il existe et fonctionne en ce moment. Puis la discussion portera sur la création des caractères phonétiques à l'écran, la caractérisation ou la spécification du système phonologique et du contexte extra-linguistique, l'algorithme de syllabation et les règles phonologiques qui ont, jusqu'à maintenant, été implantées. Enfin, les développements à venir, c'est-à-dire les ajouts et les améliorations à ce programme seront rapidement abordés.

III Objectif

PHONO est un programme qui cherche à modéliser partiellement le comportement de locuteurs typiques d'une ou de plusieurs variétés linguistiques parlées au Québec. Cette modélisation linguistique est toutefois limitée aux niveaux phonologique et phonétique du système linguistique. Concrètement **PHONO**, à partir d'une représentation phonologique donnée et d'un contexte extra-linguistique défini par l'usager, génère la représentation phonétique probable correspondante.

Ceci amène quelques remarques :

a) Il faut insister sur le fait que la modélisation contenue dans le programme **PHONO** n'est pas uniquement linguistique puisque l'application des règles phonologiques repose aussi sur une modélisation qui est faite d'un locuteur typique d'une variété linguistique.

---

[1]Communication présentée lors du 56e Congrès de l'ACFAS, Moncton, 9-13 mai 1988.

b) En ce moment, la modélisation des règles phonologiques et des contextes extra-linguistiques contenue dans **PHONO** est partielle, d'une part, parce que **PHONO** est un prototype qui ne traite qu'un nombre restreint de règles et d'autre part, parce qu'il est pratiquement impossible de formaliser toutes les règles phonologiques pour toutes les variétés utilisées dans une ou plusieurs communautés linguistiques. Il en va de même pour la modélisation des facteurs extra-linguistiques puisque la recherche n'est pas suffisamment avancée pour formuler toutes les contraintes qui tiennent compte de la position du locuteur dans la communauté linguistique ou de l'effet de la situation de communication sur la représentation phonétique.

c) *Variété linguistique* est retenu ici au sens très large. Tout «parler» commun à un groupe de locuteurs et utilisé dans un contexte extra-linguistique spécifique constitue une variété linguistique. Il n'existe à notre connaissance aucune typologie des variétés linguistiques au Québec. Concrètement, pour **PHONO**, une variété linguistique est définie en fonction de l'origine géographique et sociale du locuteur qui est actualisé. Son appartenance à tel ou tel groupe d'âge peut aussi participer à la détermination d'une variété linguistique. Jusqu'ici, **PHONO** ne tient pas compte de l'événement de communication lui-même ni des autres protagonistes.

d) La modélisation contenue dans **PHONO** se situe avant tout au niveau phonologique en ce sens que les contraintes purement phonétiques ne sont pas traitées. Par exemple, **PHONO** ne tient nullement compte de l'allongement de la voyelle devant consonne sonore et de son abrégement devant consonne sourde.

e) L'application des règles de transformation de la structure phonique est en général variable en ce sens qu'elle se fait en fonction de la probabilité d'application de chacune des règles dans un contexte linguistique et extra-linguistique donné. Par exemple : la règle d'affrication de t/d s'applique dans notre système de façon presque catégorique pour tous les locuteurs sauf pour les Beaucerons qui palatalisent les mêmes consonnes dans les mêmes contextes; inversement, l'application de la règle de diphtongaison de /wa/ est fonction de l'origine géographique du locuteur et de son appartenance à l'une ou l'autre classe sociale.

PHONO comporte, malgré ce qui vient d'être souligné, un mode **catégorique**, c'est-à-dire la possibilité d'actualiser un locuteur qui, dans son discours, réalise le plus grand nombre possible de règles et/ou les règles les plus susceptibles de s'actualiser.

IV Structure générale du programme (cf. figure 1)

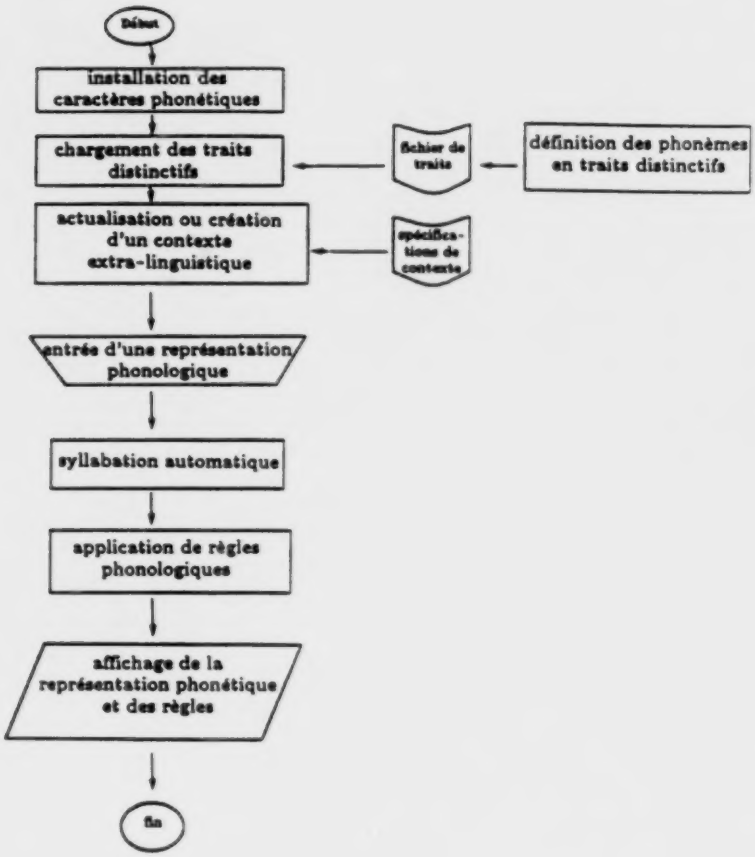
Le programme tel qu'il existe à l'heure actuelle comporte les routines et les structures de données qui apparaissent à la figure 1. Certaines des routines sont intégrées au programme **PHONO** lui-même alors que d'autres, notamment la routine d'installation des caractères phonétiques et de spécification des phonèmes en traits distinctifs, sont implantées et chargées de façon autonome.

Le programme commence par installer les symboles phonétiques qui sont nécessaires pour fournir une représentation phonologique et pour générer la représentation phonétique. Puis, une matrice de traits distinctifs pour les phonèmes utilisés est chargée en mémoire à partir du disque.

En troisième lieu, **PHONO** demande que l'usager actualise ou qu'il crée un contexte extra-linguistique. On verra plus loin en quoi consiste précisément un contexte extra-linguistique. A mesure que des contextes sont créés, ils sont sauvegardés sur disque et peuvent être rappelés ultérieurement.

FIGURE 1:

Structure générale du programme PHONO



Ces tâches préliminaires accomplies, **PHONO** demande à l'usager d'entrer une représentation phonologique. Lorsque la chaîne de symboles phonétiques est entrée, le programme syllabe automatiquement le mot et cherche à appliquer les règles phonologiques qui sont implantées. Les règles comportent des conditions ou des contraintes linguistiques et peuvent aussi comporter des contraintes sociales. Le balayage de la représentation se fait séquentiellement de gauche à droite mais peut comporter des retours arrières.

Le travail d'application des règles achevé, la représentation phonétique dérivée est affichée de même que le nom des règles qui ont été appliquées.

V Installation des symboles phonétiques

Les symboles phonétiques utilisés ont été créés et sont chargés à l'aide du programme CHARGENI. Les symboles phonétiques utilisent les positions du jeu de caractères ASCII étendu avec la carte CGA.

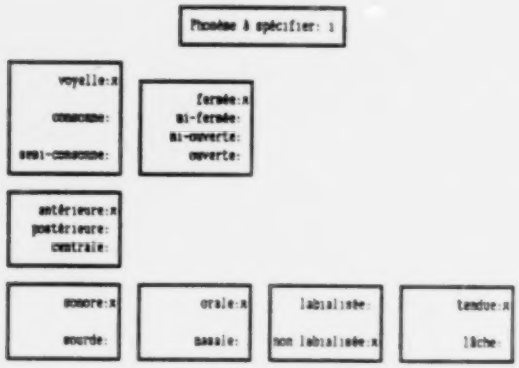
VI Chargement des traits distinctifs (cf. figure 2)

Un programme extérieur à **PHONO** permet de spécifier par une matrice de traits les phonèmes et les allophones nécessaires pour les variétés de français québécois qui sont modélisées ici. Les phonèmes sont caractérisés par une série de traits articulatoires. Une autre matrice de traits distinctifs auraient pu être choisie, mais comme le projet initial en était un de didacticiel, il a été jugé plus commode sur le plan pédagogique de retenir un vocabulaire de traits articulatoires.

Les données générées par le programme de spécification **TRAIT** sont stockées dans un fichier **PHONEME.DAT**. Ce sont ces données qui sont chargées par **PHONO** et qui sont utilisées dans l'application des règles. En effet, l'application des règles ne se fait pas sur des phonèmes spécifiques mais sur des entités partageant le même ensemble de traits distinctifs.

FIGURE 2:

Matrice de traits articulatoires

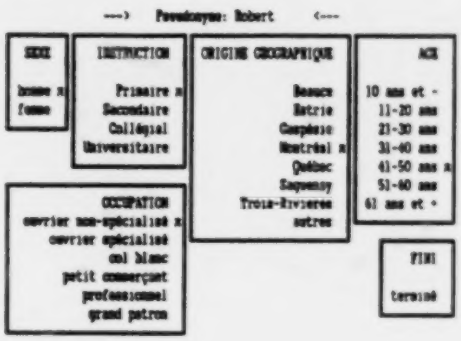


VII Spécification du contexte extra-linguistique (cf. figure 3)

La matrice de traits pertinents chargée, **PHONO** demande à l'usager s'il désire caractériser le contexte extra-linguistique dans lequel l'actualisation de la représentation phonologique prendra place. Pour le moment, ce que **PHONO** «entend» par contexte extra-linguistique est assez limité. Essentiellement, le contexte est restreint à certaines caractéristiques du locuteur : sexe, âge, instruction, et origine géographique. La figure 3 montre l'écran de définition d'un contexte. Il va sans dire que d'autres paramètres extra-linguistiques pourraient être pris en considération. Mais comme il existe peu de données précises sur l'influence, par exemple, du contexte physique ou psychologique, sur l'application de telle ou telle règle, **PHONO** se limite aux facteurs liés à la position du locuteur dans la communauté linguistique. Les contextes définis sont stockés sur disque et peuvent être rappelés en tout temps.

FIGURE 3:

Écran de définition d'un contexte extra-linguistique



L'usager a aussi le choix de refuser de définir un contexte extra-linguistique. En faisant cela, il place **PHONO** dans un mode d'application obligatoire ou catégorique des règles. En mode catégorique, **PHONO** applique le plus grand nombre possible de règles. En d'autres mots, la probabilité d'application des règles variables, à quelques exceptions près, passe à 1.

VIII Algorithme de syllabation (cf. figure 4)

Le contexte extra-linguistique défini, **PHONO** demande à l'usager de lui fournir la représentation phonologique qu'il veut voir traitée. Puis il entreprend de syllaber la chaîne de phonèmes. L'algorithme de syllabation est présenté à la figure 4.

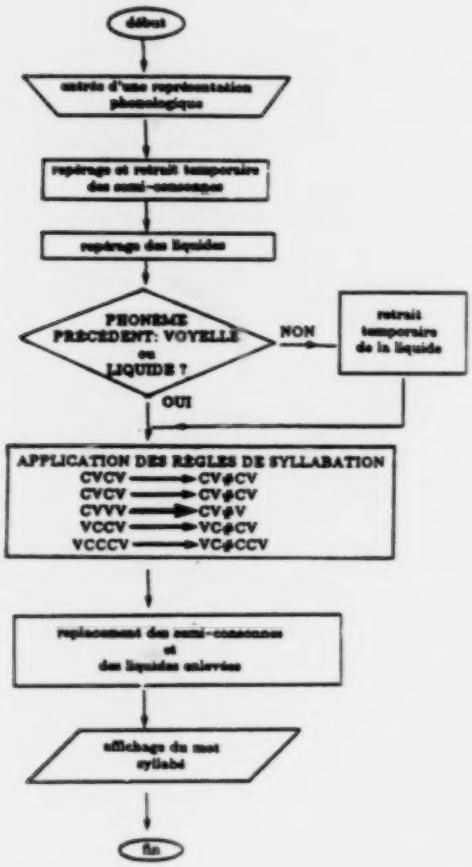
L'opération de syllabation comporte trois phases. Pour minimiser les règles possibles de syllabation, **PHONO** utilise les contraintes sur les positions possibles des semi-consonnes et des liquides dans la syllabe en français. Ainsi on sait que les semi-consonnes, à l'exception de yod dans des mots comme *paye* ou *fille*, ne peuvent apparaître que devant la voyelle de la syllabe à laquelle appartient la semi-consonne. De même la liquide qui suit une consonne qui n'est pas une liquide appartient à la même syllabe que la voyelle qui suit.

Ainsi, **PHONO** commence par retrancher les semi-consonnes et les liquides qui suivent une consonne non-liquide. Puis il cherche à appliquer autant de fois que possible une des 5 règles de syllabation. Enfin, il replace les semi-consonnes et les liquides qui ont été temporairement retranchées. Ce processus en trois étapes a l'avantage de minimiser le nombre de règles de syllabation et d'améliorer considérablement les temps de traitement.

Ainsi un mot comme *accroire* /akrwar/ est réduit en un premier temps à /akar/ puis syllabé en /a#kar/ et enfin reconstitué pour retrouver /a#krwar/.

FIGURE 4:

Algorithme de syllabation



IX Règles implantées actuellement

Les règles phonologiques implantées actuellement sont les suivantes :

VOYELLES

- allongement des voyelles longues par nature et des voyelles devant /v (vr), z, ʒ, r/ (variable)
- diphtongaison des mi-ouvertes (variable)
- diphtongaison des longues par nature (variable)
- diphtongue /wa/ (variable) --partiellement
- fermeture du /a/ postérieur en finale (catégorique)
- postériorisation du /a/ en finale (variable)
- relâchement des voyelles fermées (catégorique)

CONSONNES

- affaiblissement de la consonne /ʒ/ (catégorique)
- affrication de /t,d/ devant / i, y, j,ɥ/ (catégorique)
- aspiration des consonnes occlusives sourdes (variable)
- aspiration de /ʃ/ (catégorique)
- chute des consonnes d'un groupe consonantique final (variable)
- palatalisation de / t, d/ en Beauce (variable)
- simplification du groupe /n/ + /y/ en /r/ (catégorique)

Le programme commence par déterminer le mode articulatoire du premier segment de la chaîne. Selon qu'il s'agit d'une voyelle ou d'une consonne, il cherche ensuite à appliquer une des règles appartenant soit au groupe de règles-consonnes soit au groupe de règles-voyelles. Si aucune règle ne peut s'appliquer, le pointeur se déplace sur le deuxième segment et le travail d'examen et d'application se répète. Si une règle est susceptible de s'appliquer, cependant, **PHONO** vérifie un ensemble de conditions linguistiques ou extra-linguistiques, puis, si toutes les conditions sont rencontrées, procède à la transformation du segment. Dans le cas où la règle s'applique, le programme mémorise la position du segment transformé ainsi que le type de règle qui s'est appliquée. Selon la règle appliquée, le pointeur peut ou non se déplacer vers le segment suivant.

Prenons l'exemple de la règle d'allongement :

Telle que définie dans **PHONO**, la règle d'allongement ne peut s'appliquer que :

A : (premier groupe de conditions)

1- si la voyelle examinée n'a pas déjà été allongée, et que
2- si le segment qui suit est une consonne, et que
3- si la voyelle est une voyelle tendue (e, o, ø, ɛ̃, ɑ̃, ɔ̃, œ̃) ou une voyelle non-tendue et, dans ce dernier cas, que si la consonne qui suit est une constrictive sonore (/ v (vr), ʒ, z , r /) finale.

ou

B : (second groupe de conditions)

1- si la voyelle est tendue (e, o , ø, ɛ̃, ɑ̃, ɔ̃, œ̃), et que
2- si la voyelle est en syllabe ouverte non finale.

Si le premier groupe de conditions a été rencontré, l'allongement est catégorique. Si c'est le second groupe de conditions qui a été rencontré, l'application de la règle se fait de façon facultative. Nous avons donné arbitrairement, faute de données précises, une probabibilté de 20% d'application dans ce cas.

Un calcul de probabilité d'application est ensuite effectué. Si la règle s'applique, le segment est alors noté comme allongé.

Quand la règle d'allongement s'est appliquée, le pointeur n'est pas déplacé vers le segment suivant. **PHONO** vérifie s'il peut appliquer d'autres règles sur la voyelle allongée, notamment si la diphtongaison est possible. Chacun des segments de la représentation phonologique est ainsi traité au moins une fois.

X Développements futurs

Le nombre de règles implantées dans **PHONO** est actuellement relativement limité. Cependant l'infrastructure du programme étant déjà en place, ce sera un travail relativement facile d'ajouter d'autres règles au canevas de base. Les règles dont le domaine d'application est le mot seront d'abord considérées. Puis seront considérées les règles dont le domaine peut être extra-lexical mais dont les contraintes demeurent uniquement d'ordre phonologique. Toutes les règles qui obéissent à des contraintes d'ordre syntaxique sont évidemment exclues pour un certain temps.

Nous examinerons si l'ajout d'une interface sonore capable de produire un signal correspondant à la représentation phonétique générée par **PHONO** est de l'ordre du possible.

Les règles telles que formulées présentement ne sont pas toujours complètes ou adéquates. Il faudra donc raffiner la formulation de certaines règles.

XI Démonstration

On trouvera aux figures 5 et 6 deux exemples d'écrans générés par **PHONO**. L'usager fournit une représentation phonologique («Mot en API : »), le programme syllabe automatiquement et affiche le résultat de l'opération de syllabation, puis procède à l'application des règles en fonction du contexte extra-linguistique tel que spécifié dans le rectangle situé au coin supérieur droit de l'écran. Le résultat de l'opération est affiché dans la troisième fenêtre et le nom des règles appliquées est inscrit sous cette dernière.

FIGURE 5:

Syllabation du mot "routine"
et application de règles phonologiques

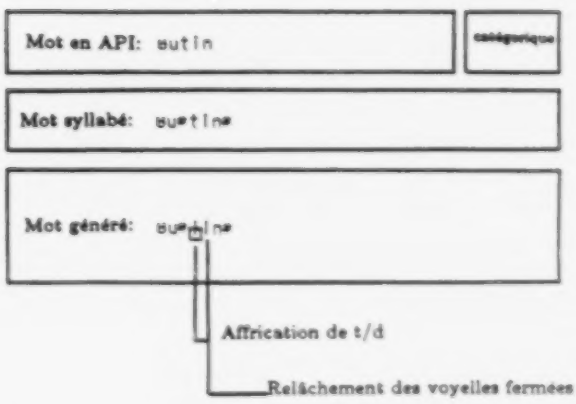
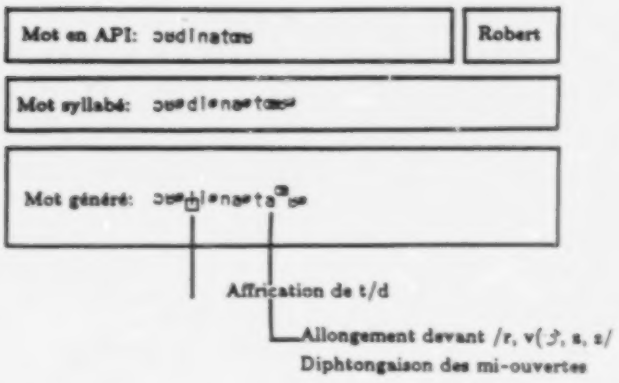


FIGURE 6:

**Syllabation du mot "ordinateur"
et application de règles phonologiques**



LA TRADUCTION ASSISTÉE PAR ORDINATEUR (TAO)[1]

Anne Bordé
Bureau des traductions
Secrétariat d'État

Introduction

Depuis plus de vingt ans déjà, le Bureau des traductions du Secrétariat d'État participe activement à des travaux de recherche et de développement en matière de traduction assistée par ordinateur (TAO).

L'intérêt du Bureau pour la TAO s'imposait de lui-même en raison de l'accroissement constant et rapide de la demande de traduction, qui est passée de 117 à 300 millions de mots entre 1967 et 1987. Le Bureau des traductions est l'un des plus grands au monde, avec un effectif de plus de 1 500 personnes. Grâce à sa banque de terminologie, **Termium III**, ses quelques 300 postes de travail, ses 25 compte-mots, son système de traduction automatique **MÉTÉO** et ses travaux de recherche et de développement, le Bureau des traductions est considéré par les communautés nationale et internationale comme étant l'un des plus avant-gardiste en matière d'intégration de services informatisés.

Il existe au Canada une très grande activité en traduction, qui répond à des impératifs politiques, culturels, sociaux et économiques. On peut estimer prudemment à quelque 600 millions de mots (4 000 ouvrages de 500 pages) le volume annuel de traduction au Canada.

Cet accroissement de la demande n'est pas exclusif au Canada. Par exemple, Van Slype et coll. estiment que le volume annuel de traduction dans le monde se situe entre 20 et 60 milliards de mots, et le coût évalué à 1,6 à 4,8 d'ECU (environ 1,7 à 5,8 millions de dollars canadiens) (Van Slype, 1981). En outre, ce volume annuel augmenterait de 9 à 10% chaque année.

Plusieurs grandes sociétés manufacturières exportatrices se sont dotées d'un important service de traduction afin de satisfaire les exigences linguistiques de leurs clients. Pour les entreprises de haute technologie, cela n'est pas une tâche facile. Par exemple, la documentation relative à un élément d'un équipement complexe (par exemple, un ordinateur, une génératrice d'électricité, etc.) peut comporter des millions ou des dizaines de millions de mots. En fait, la documentation relative aux équipements complexes représente toujours une part importante des coûts de production de tels produits. Selon Brinkman, de la multinationale allemande Siemens, cette part atteint, dans certains cas, 50% du prix de revient d'un produit (Brinkman 1979).

1. Phase I (1965 - 1970)

Le Centre d'études pour le traitement automatique de données linguistiques (**CETADOL**) est créé à l'université de Montréal sous la direction de Guy Rondeau. Les études et recherches préliminaires de ce centre sont orientées vers la conception d'un système de traduction automatique de deuxième génération.

---

[1]Communication présentée lors de la Deuxième Conférence des chefs d'État et de Gouvernement de pays ayant en commun l'usage du français, sous le thème : Industrie de la langue, Québec, 5 septembre 1987.

Ainsi les recherches portent principalement sur les nouvelles théories linguistiques (ex. la grammaire générative - transformationnelle), la théorie des automates et des grammaires formelles, l'expérimentation des principaux algorithmes d'analyse syntaxique existants (ex. les algorithmes **Cocke, Kuno-Woods, Sakai**, etc.), la mise au point de nouveaux formalismes informatiques (ex. les W-grammaires), la formalisation des structures de représentation de l'information et, enfin, la description et la formalisation de divers phénomènes linguistiques (ex. en morphologie, en syntaxe et en sémantique).

Durant cette première phase jusque vers 1972, le Conseil national de recherches du Canada finance la recherche fondamentale.

2. Phase II (1970 - 1976)

Au début de 1970, le **CETADOL** précise son orientation et devient le Groupe de recherche en *t*raduction *a*utomatique de l'*u*niversité de *M*ontréal (**TAUM**) sous la direction successive de Alain Colmerauer et de Richard Kittredge.

On s'engage dans la conception, le développement et la mise au point du premier prototype d'un système de traduction automatique de deuxième génération. Ce premier prototype s'appelle **TAUM-71**. Suivent les prototypes **TAUM-73**, **TAUM-75** et **TAUM-76**.

Concurremment, on travaille à la conception, au développement et à l'implantation de divers langages de programmation de haut niveau (ex. **Systèmes-Q**, **TARZAN**, **REZO**).

A partir de 1973, le Bureau des traductions du Secrétariat d'État prend en main la poursuite de la recherche et du développement de systèmes.

En 1975-1976, **TAUM** fait la conception, le développement et la mise au point du système **TAUM-MÉTÉO** pour le compte du Bureau des traductions. En 1976, **TAUM** livre le prototype **TAUM-MÉTÉO** au Centre météorologique canadien (CMC). Le système **MÉTÉO**, comme on le nomme aujourd'hui, demeure à ce jour le seul système de traduction automatique de deuxième génération au monde pouvant fonctionner jour et nuit, 365 jours par année.

3. Phase III (1976 - 1980)

En 1976, le groupe **TAUM** et le Bureau entreprennent de relever un défi de taille : la conception, le développement et la mise au point d'un système de traduction automatique de deuxième génération en vue de traduire les manuels d'entretien de l'avion **AURORA (CP-140)**, le nouveau patrouilleur maritime du ministère de la Défense nationale.

En 1979, **TAUM** fait une démonstration du prototype **TAUM-AVIATION**. Si le système **TAUM-AVIATION** se révèle alors être en mesure d'aider à la traduction des manuels d'entretien des circuits hydrauliques, il ne pourra vraisemblablement pas servir à traduire les autres manuels, à moins d'autres importants investissements tant en matière de développement que de recherche. Une évaluation des coûts-avantages s'impose donc. Toutefois, les progrès techniques et scientifiques accomplis depuis la conception du système **MÉTÉO** sont très importants, car **TAUM** a mis au point une technologie de deuxième génération qui se range parmi les plus avancées au monde.

4. Phase IV (1980 - 1983)

4.1 Évaluation

En 1980-1981, le Bureau procède à une évaluation des coûts-avantges du système **TAUM-AVIATION**. On tente alors d'évaluer la performance linguistique du système et d'en analyser les coûts de développement et d'exploitation.

L'évaluation conclut d'abord que le système donne des résultats satisfaisants quand il produit une traduction. Par exemple, la cote générale attribuée à la portion traduite de l'échantillon test, des points de vue de l'intelligibilité, de la fidélité et du style, représente environ 80% de la traduction humaine. De plus, on note que les coûts directs de la traduction automatique révisée pourraient devenir inférieurs à ceux de la traduction humaine. Toutefois, le coût de création des dictionnaires et de maintien d'une équipe d'exploitation a un impact négatif sur la rentabilité éventuelle de la traduction automatique des manuels du **CP-140** à l'aide du système **TAUM-AVIATION**. Enfin, pour rentabiliser ce système, il faut que deux conditions soient remplies simultanément : la réduction des coûts directs de la traduction automatique et la traduction au moyen du système de 5 à 6 millions de mots par année. A la lumière de cette évaluation, on estime qu'il est impossible d'affirmer que le système **TAUM-AVIATION** peut, dans un avenir prévisible, se révéler plus rentable que la traduction humaine en raison surtout des coûts indirects (ex. fabrication des dictionnaires et exploitation du système).

4.2 Étude de la faisabilité de la traduction automatique

En 1981, le Bureau effectue une étude de faisabilité sur la traduction automatique. Les principaux objectifs de cette étude consistent à mesurer les possibilités d'extension du système **TAUM-AVIATION** aux manuels d'électronique du **CP-140**, à rechercher les textes confiés au Bureau qui se prêteraient le mieux à la traduction par ordinateur, à évaluer divers systèmes commerciaux de traduction et à produire un rapport recommandant des priorités et un plan d'implantation, s'il y a lieu. Les conclusions de cette étude sont les suivantes :

a) Le système **TAUM-AVIATION** peut être appliqué aux manuels d'électronique du **CP-140**. De plus, ce système est en mesure d'en produire des traductions comparables à celles obtenues dans le domaine de l'hydraulique. Toutefois, le coût de fabrication des dictionnaires d'électronique demeure l'inconvénient principal.

b) Les avis de concours, les descriptions de fonctions et le *Journal des marques de commerce* semblent satisfaire aux critères de volume, de répétition ou d'urgence alors retenus pour choisir les domaines prioritaires d'application de la traduction automatique. Toutefois, cette étude n'a pas permis d'établir des critères scientifiques permettant d'évaluer la complexité linguistique d'un texte afin de déterminer s'il était possible de le traduire par ordinateur.

c) Parmi les systèmes commerciaux étudiés, **SYSTRAN II**, **WEIDNER** et **ALPS** semblent les plus prometteurs. Toutefois, avant de recommander ou, encore, d'envisager l'implantation d'un de ces systèmes, il est suggéré d'effectuer un certain nombre d'essais.

4.3 Évaluation et essai de systèmes commerciaux

En 1981-1982, le Bureau fait un essai et une évaluation du système **ALPS**. A la suite de cet essai, on conclut que le système **ALPS** présente des faiblesses en matière d'analyse de la langue-source et de génération de la langue d'arrivée et qu'il n'offre donc pas le niveau d'aide requise pour la traduction des avis de concours, des descriptions de fonctions et du *Journal des marques de commerce*.

En 1982, le Bureau lance une demande de propositions afin de solliciter des offres de systèmes de traduction automatisée aptes à aider à la traduction des avis de concours, du *Journal des marques de commerce* et des descriptions de fonctions. Seul le système **Weidner** est retenu et, après les épreuves, ne se révèle pas supérieur au système **ALPS**.

4.4 Autres études

En 1983, le Bureau entreprend l'élaboration d'une méthodologie générale d'évaluation linguistique de systèmes de traduction automatisée. L'objet de cette méthodologie est de définir et d'expliquer en détail les principaux éléments à évaluer avant de procéder à l'essai ou à l'achat d'un système de traduction automatisée.

En 1983 également, le Bureau subventionne une étude portant sur l'élaboration de critères scientifiques pour déterminer la complexité linguistique des textes à traduire et, ainsi, les classer par ordre de difficulté croissante. Des échantillons de 17 types de textes que traduit le Bureau servent de matière à cette étude qui révèle que certains types de textes sont à la portée de la technologie de deuxième génération comme **MÉTÉO** ou **TAUM-AVIATION**, mais leur très faible volume ne justifie pas le développement immédiat de systèmes. D'autre part, ils sont tous, sans exception, hors de la portée des systèmes moins évolués sur le plan de l'analyse de la langue-source.

5. Phase V (1984 - 1987)

5.1 Relance de la recherche et du développement au Canada

Le Bureau s'emploie activement à promouvoir la recherche et le développement au Canada en matière de **TAO**. Le Bureau, toujours convaincu de la justesse, de l'originalité et de l'avancement technique et scientifique des travaux canadiens en **TAO**, sollicite divers appuis au sein du gouvernement canadien afin de mieux encadrer les investissements qui pourraient être faits dans ce domaine.

C'est ainsi qu'en juillet 1983, le Secrétariat d'État et le ministère des Communications se concertaient pour confier à la société **Quasar Systems Ltd**, maintenant la **Cognos Corporation**, une étude sur le traitement automatisé des langues naturelles et sur l'application de l'intelligence artificielle à ce domaine.

Le rapport **Cognos**, présenté en 1984, recommandait trois fronts sur lesquels la recherche et le développement pouvaient être poursuivis et entrepris concurremment. Ils constituent maintenant les objectifs qui sous-tendent la collaboration entre le ministère des Communications et le Secrétariat d'État :

a) la mise au point d'un poste de travail professionnel dans le contexte de la bureautique, à titre de projet à court terme;
b) le raffinement de la technologie de traduction automatique de deuxième génération, telle qu'elle est représentée par le prototype **TAUM-AVIATION**, et son application à des textes plus compliqués que les bulletins météorologiques, mais moins que les manuels d'entretien d'avions, textes qui offriraient un volume suffisant et se situeraient dans un sous-langage défini, à titre de projet à moyen terme;
c) la recherche et le développement d'une technologie avancée de traduction automatique ayant recours à des techniques d'intelligence artificielle et de traitement des langues naturelles.

5.2 Poste de travail du traducteur

Le projet de développement d'un poste de travail pour traducteur canadien a été confié au Centre de recherche en informatisation du travail (CCRIT). A l'été 1985, le CCRIT commandait à la firme Socioscope Inc. une étude visant à définir les caractéristiques fonctionnelles du poste. Le rapport, déposé en 1985, proposait d'automatiser toute une gamme de fonctions qui exigeaient parfois le recours à des techniques très avancées et très coûteuses, mais il ne les classait pas par ordre de priorité et n'examinait pas les produits disponibles sur le marché qui auraient pu être adoptés à court terme. Le CCRIT a donc repris les conclusions du rapport afin de faire ressortir les composantes les plus importantes et les plus rentables à inclure. Il est prévu qu'à l'automne de 1988, un prototype pourra être mis à l'essai dans un certain nombre de services de traduction.

5.3 Études de marché

Afin de démontrer aux chercheurs et au monde des affaires la viabilité économique du créneau de la TAO et de les inciter à investir temps et argent, le ministère des Communications et le Secrétariat d'État confiaient en janvier 1987 à la firme Coopers & Lybrand Consulting Group l'exécution d'une étude de marché portant sur les facteurs suivants : la demande et l'offre en système de TAO, la capacité de l'industrie canadienne à faire face à la demande, sa compétitivité par rapport aux autres pays et les possiblités d'investissement. Le rapport définitf, qui présentera des recommandations quant aux mesures à prendre, est prévu pour septembre 1987.

5.4 Évaluation de systèmes de TAO

Le Bureau des traductions continue d'évaluer les systèmes commerciaux d'application générale.

En 1985, il met à l'essai le système MicroCat de Weidner sur quatre types de textes choisis pour leur style assez égal. Au bout de quatre mois, la conclusion est que le système n'est pas rentable parce que la mauvaise qualité des traductions brutes exige trop de post-édition.

En mars 1987, un projet d'un an est lancé dans les services de traduction à Montréal, avec la société Logos. Pendant cette période, trois traducteurs du Bureau et trois experts du constructeur utilisent trois versions successives du logiciel afin de traduire de façon opérationnelle des textes d'un gros volume dans le domaine de l'informatique (deux autres domaines pourront être envisagés ultérieurement) et de recommander des améliorations au constructeur. Au terme de l'expérience, le Bureau aura des données précises sur la rentabilité du système et disposera, vraisemblablement, d'un système taillé sur mesure pour répondre à une certaine demande.

5.5 Programme national de recherche et de développement

Le ministère des Communications a préparé un mémoire au Cabinet, approuvé en juin 1987, destiné à la mise en place d'une stratégie canadienne sur les techniques de l'information et les communications. Cette stratégie sous-tendra un futur proramme national de recherche et de développement en TAO.

BIBLIOGRAPHIE

VAN SLYPE et coll., *Mieux traduire pour mieux communiquer*, Infotecture pour la Commission des communautés européennes, Luxembourg, 1981, 174 p.

BRINKMAN, (K.-H.), "Perspectives d'avenir de la traduction automatique", dans *META*, vol. 24, no 3, 1979.

DIX ANS DE MÉTÉO[1]

John Chandioux
Experts-conseils Inc.

Introduction

MÉTÉO est un système de traduction automatique anglais-français exploité par le Bureau des traductions du Secrétariat d'État pour le compte d'Environnement Canada. Le système traduit les prévisions météorologiques destinées au grand public émises pour l'ensemble du Canada et fonctionne 24 heures par jour. La charge du système est supérieure à 30 000 mots par jour; le taux de révision est de l'ordre de 5% et le prix de revient du mot-machine brut est inférieur à un demi cent. Plus étonnant encore, le premier système **MÉTÉO** a été implanté voici plus de dix ans et on le cite encore aujourd'hui comme l'un des rares succès de la traduction automatique.

Un peu d'histoire

Le prototype du premier système **MÉTÉO** a été développé par le groupe **TAUM** de l'université de Montréal en 1975-1976 dans un langage appelé **Systèmes-Q**.

Après avoir été mis au rancart pendant un an, la bande magnétique du système **TAUM-MÉTÉO** a été confiée par le Bureau des traductions à la société **Consultants en linguistique computationnelle limitée**, qui a développé la première version commerciale, connue sous le nom de **MÉTÉO 1**. Le passage du prototype à l'exploitation ne s'est pas fait sans surprise. Le système **TAUM-MÉTÉO** ne traduisait effectivement que 40% des phrases rencontrées et il a fallu une année de développement supplémentaire pour que le système **MÉTÉO 1** atteigne le taux de 80% visé au départ.

En 1982, la société **John Chandioux experts-conseils Inc.** a terminé le développement d'un nouveau langage de programmation, appelé **GramR**, et a été chargé par le Bureau des traductions d'étudier la faisabilité d'une traduction automatique sur micro-ordinateur. Peu après, la société commençait à développer de sa propre initiative un nouveau système, baptisé **MÉTÉO 2**.

En 1983, le Bureau des traductions a accordé à la société **John Chandioux experts-conseils Inc.** un contrat pour un essai comparé de trois mois des systèmes **MÉTÉO 1** et **MÉTÉO 2**. Le résultat de l'essai a révélé que le système **MÉTÉO 2** est plus performant, plus fiable et plus convivial que son prédécesseur.

Depuis le 10 octobre 1984, le Bureau des traductions loue le système **MÉTÉO 2** avec une formule «clé en main» qui comprend les ordinateurs, les logiciels, le service 24 heures par jour et la mise à jour des dictionnaires selon les besoins.

---

[1]Communication présentée lors de la Deuxième Conférence des chefs d'État et de Gouvernement de pays ayant en commun l'usage du français, sous le thème : Industries de la langue, Québec, 5 septembre 1987.

Un conflit de générations

Commençons par situer le système **MÉTÉO** dans l'histoire de la traduction automatique. Au départ, la recherche est un reliquat de la guerre froide et a été subventionnée massivement par les militaires vers la fin des années 50, d'abord aux États-Unis, puis en France. Le problème de la traduction automatique était alors assimilé à un simple problème de décryptage donnant naissance aux systèmes dits de première génération. Ceux-ci se composent d'un gigantesque dictionnaire auquel on associe une analyse du contexte, généralement limitée aux deux mots qui suivent et qui précèdent le mot courant. On peut ainsi trancher certaines ambiguïtés grammaticales.

Le plus connu de ces systèmes est le système **SYSTRAN** utilisé par la **Foreign Technology Division** de l'Armée de l'air américaine. Le système donne une traduction sommaire de tous les textes scientifiques en langue russe que le service peut obtenir. Les traduction brutes sont ensuite examinées par des spécialistes qui identifient les textes ayant un intérêt stratégique et ceux susceptibles d'intéresser la communauté scientifique. Les quelques textes rètenus sont en général retraduits manuellement.

En 1966, un rapport de la **National Science Foundation** concluait à l'impossibilité de la traduction automatique et donnait un coup de grâce à la recherche. La publication de ce rapport ne tenait pas compte de l'émergence d'une deuxième génération de systèmes.

Le système de deuxième génération possède trois caractéristiques fondamentales :

1- il est programmé dans un langage évolué, directement accessible au linguiste;
2- il analyse la totalité de chaque phrase;
3- il effectue la traduction en trois phases appelées **ANALYSE**, **TRANSFERT** et **GÉNÉRATION**.

Cette dernière distinction correspond aux trois étapes fondamentales de l'opération de traduction : la compréhension du texte-source, la transposition de l'idée et la formulation de l'idée résultante dans la langue-cible.

Le système **MÉTÉO** est la première application commerciale de la deuxième génération et serait, si l'on en croit certains spécialistes, la seule à ce jour.

Depuis, la troisième génération de traduction automatique a vu le jour dans les laboratoires de recherche. Le système de troisième génération associe au système de deuxième génération une base de connaissances dans laquelle le système peut puiser pour trancher les ambiguités de sens.

Le langage GramR (MD)

La partie traduction du système **MÉTÉO 2** est écrite dans un métalangage appelé **GramR**. Techniquement parlant, il s'agit d'un transducteur de chaînes d'arborescences paramétrées. Pratiquement parlant, il permet au linguiste, familier avec le formalisme de la grammaire transformationnelle développée par Chomsky, d'écrire son modèle linguistique sous forme de règles, de le compiler et de l'évaluer sur des exemples de son choix.

Le développement du langage **GramR** a été fortement influencé par les travaux du Groupe d'Études pour la Traduction Automatique de Grenoble et nous sommes particulièrement redevables au regretté professeur Bernard Vauquois pour ses nombreuses suggestions et ses encouragements.

En 1982, **GramR** était le tout premier langage de deuxième génération sur micro-ordinateur; aujourd'hui il a fait ses preuves dans un grand nombre d'applications : la traduction automatique (**MÉTÉO 2**), la détection de fautes d'orthographe (**SpellR**), l'accentuation automatique (**Accents**), l'enseignement du français (**Le Groupe Nominal**) et les interfaces en langue naturelle (**EASY-DOS**).

Enfin, le langage **GramR** est désormais disponible en versions MS-DOS, UNIX 68000 et VAX VMS.

Les grammaires

Les grammaires du système **MÉTÉO 2** effectuent une analyse complète de chaque phrase avant la traduction. On pourrait penser que le fait de travailler sur des textes en langage télégraphique faciliterait la tâche : il n'en est rien. Ce type de texte est entièrement en majuscules, pratiquement sans ponctuation et les mots sur lesquels reposent une analyse automatique conventionnelle, tels que les conjonctions, les articles et les prépositions, sont souvent absents. Pour cette raison, l'analyseur du système **MÉTÉO 2** est beaucoup plus sémantique que celui utilisé dans le système prototype pour la traduction de manuels d'informatique.

Fort heureusement, la sémantique d'un micro-monde comme la météorologie est facile à circonscrire. Un énoncé bien formé se résume en général à une condition météorologique, souvent limitée dans le temps ou dans l'espace. Cette condition pouvant débuter, se maintenir, cesser ou faire place à une autre.

Les quinze grammaires du système **MÉTÉO 2** se divisent en cinq types fondamentaux :

1- **La pré-édition.** Ces grammaires ont pour objet d'éliminer les redondances ou les mots non significatifs, de normaliser certaines formulations, d'identifier les expressions idiomatiques et de développer les abréviations.
2- **Le dictionnaire.** Le dictionnaire associe à chaque adjectif, nom, ou verbe, la catégorie syntaxique et certaines informations sémantiques et morphologiques utilisées par les grammaires suivantes.
3- **L'analyseur.** L'analyseur identifie d'abord les groupes nominaux et les ventile en conditions météorologiques, en circonstanciels de temps et en circonstanciels de lieu. Cette partie de l'analyse repose sur les compatibilités et incompatibilités de classes sémantiques. Puis vient l'analyse du groupe verbal et sa combinaison avec les groupes nominaux en énoncés bien formés.
4- **La génération syntaxique.** Ces grammaires opèrent une série de transformations sur le résultat de l'analyse afin de générer les mots de la phrase-cible dans le bon ordre.
5- **La génération morphologique.** Ces grammaires traitent les problèmes d'accord en genre et en nombre, d'élision, de contraction et effectuent d'ultimes ajustements stylistiques.

On remarquera que la phase de transfert de la définition du système de deuxième génération semble avoir disparue. En fait, le transfert est effectué au niveau du dictionnaire et surtout de l'analyseur. Ceci découle en partie du fait que l'analyse est plus sémantique que syntaxique.

L'environnement MÉTÉO 2

Le système **MÉTÉO 2** ne se limite pas à une série de grammaires en langage **GramR**; il s'agit d'un environnement complet qui totalise plus de 10 000 lignes de programmation. Le cheminement d'un bulletin est le suivant :

GERTELEX, le logiciel de communication, lit la ligne TÉLEX, isole les bulletins à traduire, prend note de l'indicatif du bulletin, du nombre de caractères et de l'heure d'arrivée, puis met le bulletin en file d'attente de révision.

EXEMÉTÉO, une version personnalisée de l'exécuteur **GramR**, prend le bulletin en charge, compte les mots, applique les grammaires de traduction, prend note du nombre de marques de révision générées par l'analyseur et place le bulletin en file d'attente de révision.

SUPRVIZR, le moniteur, affiche en tout temps l'état des files d'attente de traduction de révision et de transmission. S'il y a des bulletins en file d'attente de révision, le traducteur spécifie le bulletin qu'il veut réviser.

RÉVISEUR, un éditeur spécifiquement conçu pour la traduction automatique, permet au traducteur de réviser le bulletin en un minimum d'opérations. Il compte également le nombre de corrections effectuées, demande les initiales du réviseur, met le texte en format TÉLEX, prend note de l'heure de fin de révision et met le bulletin en file d'attente de transmission.

Il y a également une batterie d'utilitaires pour relancer automatiquement le système après une coupure de courant ou une panne, effectuer un diagnostic complet après la relance, pour charger de nouvelles grammaires sans interruption du fonctionnement et fournir au chef de section des rapports de production et des statistiques de fonctionnement.

Le matériel

L'ordinateur utilisé est un micro-multi-usagers de marque **Cromemco** centré sur un microprocesseur 68010. Il est doté d'un disque rigide de 5 Mo et d'une mémoire centrale de 512 Ko. Une interface de communication série multi-ports assure la liaison avec une ligne TÉLEX à 300 bauds du CN/CP et un terminal intelligent servant de poste de révision.

Le système d'exploitation est une variante d'UNIX appelée CROMIX.

Le système a été choisi pour sa performance et sa fiabilité. Le temps de panne est présentement inférieur à une journée par an et un second ordinateur est prêt à prendre la relève en cas de besoin. La plus longue interruption de service à ce jour est due, ironie du sort, à la météo. Le service de traduction a été interrompu pendant 64 heures consécutives à la suite du violent orage qui a frappé Montréal le 14 juillet 1987.

IMPLANTATION DU SYSTÈME DE TRADUCTION ASSISTÉE PAR ORDINATEUR (TAO)[1]

Moïse Lévy
Bureau des traductions
Secrétariat d'État du Canada

Ce printemps, le secteur **Langues officielles et traduction** du Secrétariat d'État lançait à Montréal un projet de traduction assistée par ordinateur. Ce projet devant se dérouler en milieu de travail, c'est la section **Informatique** qui a été choisie pour des raisons à la fois stratégiques et techniques. En effet, le volume de la demande dans cette spécialité ainsi que le bassin de connaissances techniques acquises dans cette section semblaient justifier le choix du ministère.

Un calendrier d'implantation et de mise en oeuvre, couvrant la période d'essai initiale d'un an, a été établi d'un commun accord entre les représentants du Secrétariat d'État et ceux de la société **Logos**, dont nous allions utiliser le système.

L'équipe de traduction affectée à ce projet avait pour tâche d'enrichir la base de données terminologiques du système, d'élaborer des méthodes de travail susceptibles d'en accroître l'efficacité et d'examiner les qualités ergonomiques de ce nouvel outil de travail. Elle devait, en outre, évaluer la performance du système en fonction de trois versions successives du logiciel de traduction.

Phases du projet

Dans l'ensemble, le projet devait comporter quatre phases distinctes : le choix et la formation du personnel, l'adaptation du personnel au système, la production et enfin, l'évaluation.

1. *Choix et formation du personnel* : dans un milieu de travail hautement technologique, nous ne pouvions choisir que des traducteurs reconnus pour leur compétence professionnelle et leur autonomie en milieu de travail. De plus, il fallait choisir parmi ceux que la présence d'un ordinateur dans nos locaux n'effrayait pas. Trois traducteurs de la section **Informatique** ont été choisis pour suivre des cours d'introduction à l'équipement et au traitement de texte **Wang** ainsi qu'au système **Logos**. Trois autres traducteurs, travaillant pour la société **Logos**, sont venus se joindre à l'équipe.
2. *Adaptation du personnel au système* : la durée d'apprentissage et de familiarisation aux différentes fonctions du système a été évaluée à environ un mois. Pendant cette période, les traducteurs ont effectué la traduction de textes d'exercices et d'un corpus de référence composé de 200 phrases-type extraites des textes de nos principaux clients, ont enrichi le dictionnaire et ont établi une liste de recommandations destinées au personnel chargé de développer le système.
   Dès le début, il a fallu se doter d'une méthode et d'outils mieux adaptés aux besoins particuliers du Bureau.
3. *La production* : pour mieux mesurer l'adaptabilité du système à notre spécialité, nous avons décidé d'offrir un service efficace à la clientèle. Ainsi, tout en essayant d'enrichir et d'améliorer le système, nous devions nous efforcer de le rendre rentable en milieu de travail.

[1]Communication présentée lors de la Deuxième Conférence des chefs d'État et de Gouvernement de pays ayant en commun l'usage du français, sous le thème : Industries de la langue, Québec, 5 septembre 1987.

4. *L'évaluation* : outre l'évaluation du système comme outil de traduction, appelé à évoluer par l'introduction de trois versions successives du logiciel, il fallait également juger son efficacité sur le plan de la production, de la qualité des textes produits et des aspects ergonomiques qui le caractérisaient.

Choix des textes

Pour le moment, les textes destinés à être traduits par ordinateur doivent répondre à certains critères de base : textes volumineux, comme les guides et manuels d'exploitation ou de formation; textes suivis ne comportant pas beaucoup de tableaux, de colonnes ou de chiffres; textes ayant des parties répétitives ou se recoupant d'un chapitre ou d'un volume à l'autre; textes dont la rédaction se prête à une assimilation plus directe par la machine.

La rentabilité du système repose en partie sur la réception du texte de départ sous forme «ordinologique» (sur disquette ou par télécommunications) afin d'en éviter la saisie manuelle. Un sérieux effort d'information a donc été entrepris auprès des services ministériels de traduction afin qu'ils renseignent les clients sur l'importance de remettre les disquettes du texte de départ, et si possible des disquettes d'un format acceptable.

La traduction assistée par ordinateur exige du traducteur qu'il franchisse plusieurs étapes précises et de durée variable selon la nature et le degré de technicité du texte de départ. Avant de soumettre un texte à la traduction, il doit établir un plan de travail au terme duquel il aura reconnu les principales difficultés que le système **Logos** devra résoudre, en même temps qu'il aura déterminé les domaines spécialisés à spécifier au système. C'est là une tâche primordiale car la qualité des résultats obtenus dépend en grande partie des «codes de ministère» et des «codes de domaines» que le dictionnaire ira consulter en priorité, avant de tous les passer en revue, le cas échéant. Le traducteur procède ensuite à une pré-édition du texte de départ afin d'éliminer toutes sortes d'erreurs comme les fautes d'orthographe ou les sigles erronés, erreurs qu'il ira corriger dans le texte de départ. La pré-édition permet donc d'alléger et d'accélérer les tâches ultérieures.

Le traducteur entreprend ensuite deux tâches successives : la recherche, par le système, des termes nouveaux (New Word Search) et la détermination, par le système, de la fréquence des expressions nominales (Noun Phrase Search), chacune de ces tâches étant elle-même suivie d'une séance d'enrichissement du dictionnaire (fonction ALEX). C'est à ce stade que le système devient interactif et que l'intervention du traducteur devient indispensable. En fonction des résultats recherchés, le traducteur doit donner une série de précisions grammaticales ou syntaxiques au sujet de chaque nouveau terme ou de chaque expression nouvelle qu'il introduit dans le système.

Après une analyse des résultats, à l'issue de la première boucle ou série de tâches, le traducteur peut recommencer le même processus pour s'assurer qu'il a pleinement utilisé les ressources du système. Il peut aussi rédiger de nouvelles règles sémantiques (fonction SEMANTHA) qui permettent au système de traduction de considérer le contexte sémantique et syntaxique et de choisir un transfert plus indiqué pour un mot ou une locution dans la langue de départ. Les règles sémantiques permettent d'identifier les contextes dans lesquels les tranferts de la base du dictionnaire seraient maladroits ou inappropriés et remplacent ces transferts par des constructions qui donnent une signification plus exacte dans la langue d'arrivée. Le traducteur doit enfin procéder à la post-édition du texte en vue de la livraison au client.

Un grand nombre de tâches se déroulant en arrière-plan, le traducteur doit modifier ses méthodes de travail et s'adapter au nouvel outil qui lui est offert. Très souvent, il passera la plus grande partie de ses heures de travail à accomplir des tâches d'analyse, de reformatage, de pré-édition et d'enrichissement du dictionnaire et n'entreprendra sa traduction qu'en fin de journée; la traduction ne

sera prête que le lendemain. C'est à ce moment qu'il pourra analyser les résultats et procéder à la post-édition de son texte. L'interface homme-machine est donc un élément indispensable de la traduction assistée par ordinateur.

Par conséquent, nous sommes confrontés à un changement radical du travail. Le rôle du traducteur se trouve totalement bouleversé. Non seulement doit-il s'éloigner du schéma traditionnel de la traduction - assimilation, conversion et rédaction - mais encore doit-il changer son rythme de travail pour s'adapter au fonctionnement et aux exigences de la machine. Certaines opérations lui apparaissent lourdes et encombrantes alors que d'autres, le transfert par exemple, ne lui imposent plus cette fatigue mentale, grande génératrice de stress, qui caractérise la profession de traducteur.

Au stade de la post-édition, les nouvelles fonctions du traducteur ressemblent davantage à celles du réviseur traditionnel. Désormais, le traducteur doit s'efforcer de corriger sur les sorties-machine ce qu'il n'a pu trouver ni introduire dans le programme de traduction.

Nous pouvons, d'ores et déjà, imaginer qui sera le traducteur de demain. Devant un appareil de post-édition hautement perfectionné qui lui permettra d'accomplir toutes les tâches préparatoires, il pourra aussi consulter simultanément le texte de départ et le texte d'arrivée, et accéder aux bases de données d'un dictionnaire électronique comme **Termium III**. Il n'aura plus à manipuler du papier ni à consulter constamment les volumineux dictionnaires et encyclopédies qui ont, jusque-là, constitué les piliers indispensables de sa profession.

LA BANQUE DE TERMINOLOGIE DU QUÉBEC: UN INSTRUMENT DE DÉVELOPPEMENT SOCIAL, AU CARREFOUR DES ENJEUX TECHNOLOGIQUES ET LINGUISTIQUES[1]

Jean-Marie Fortin
Directeur des productions linguistiques et terminologiques
Office de la langue française

La Banque de terminologie du Québec : un instrument de développement social, au carrefour des enjeux technologiques et linguistiques, voilà un titre évocateur des nombreuses transformations sociales qu'a connues le Québec au cours des deux dernières décennies, notamment à la faveur des législations linguistiques. Titre tout aussi évocateur des changements technologiques que les divers agents de la société sont en train de réaliser et qui bouleverseront le fonctionnement de nos vies au travail, à l'école et à la maison.

C'est donc en tant que témoin de l'usage et agent de changement que la **Banque de terminologie du Québec**, qu'on appelle aussi de façon plus familière la **«BTQ»**, vous est présentée aujourd'hui.

Le cadre historique de la situation linguistique

Auparavant, il est nécessaire de se pencher sur l'évolution historique du Québec, dans ce qu'il a de plus significatif, sur le plan linguistique. Les vingt-cinq dernières années ont été marquées dans tout le monde occidental par une extension phénoménale de la capacité d'intervention de l'État dans divers secteurs de l'activité humaine. Les nombreuses manifestations de la réglementation sociale que l'État prescrit visent habituellement à amener «les diverses composantes de la société à assumer des tâches et des responsabilités qui ne leur sont pas dictées par les incitations normales du marché» (Allaire et Miller 1980:5), que ce soit sur le plan de la protection du consommateur, sur le plan de l'environnement ou de la sécurité au travail. Les questions linguistiques n'ont pas échappé à cette tendance. En Norvège, en Turquie, en Israël, au Canada et au Québec, et à bien d'autres endroits encore, l'État est intervenu pour établir l'équilibre souhaité par la population dont la conscience linguistique n'a jamais été aussi aigüe qu'en cette seconde moitié du XX[e] siècle. «La langue peut donc jouer un rôle mobilisateur extrêmement puissant dans l'intégration ou la désintégration des sociétés» (Martin 1983:10), ainsi que nous l'avons vécu au Québec et au Canada au cours des deux dernières décennies.

La situation linguistique québécoise a ceci de particulier, qu'on a voulu, à la faveur des législations linguistiques successives, corriger rapidement une situation qui résultait de l'industrialisation qui s'est faite en anglais au XIX[e] siècle et dans la première moitié du XX[e] siècle, tout en continuant à progresser dans les diverses manifestations industrielles, économiques et sociales.

---

[1]Communication présentée lors de la Deuxième Conférence des chefs d'État et de Gouvernement de pays ayant en commun l'usage du français, sous le thème : Industries de la langue, Québec, 5 septembre 1987.

La BTQ : un instrument de développement social

Sur le plan de la langue elle-même, il a donc fallu faire un rattrapage historique bien particulier au Québec, compte tenu de la situation dans les autres pays de la francophonie, où la terminologie traditionnelle des domaines d'activités industrielles et économiques était connue et utilisée, tandis qu'au Québec l'anglais dominait dans la plupart des secteurs d'activité. Il fallait donc trouver le moyen de substituer les termes techniques français aux termes techniques anglais en usage principalement dans le milieu des entreprises et des affaires. Pour faire cette substitution, il fallait que les ressources terminologiques soient disponibles et facilement accessibles. La **Banque de terminologie** est donc née de la nécessité. Elle a été créée en 1973 au coeur même du débat linguistique. En raison de la situation linguistique particulière du Québec par rapport au reste de la francophonie, il a fallu dans un premier temps procéder à des inventaires terminologiques dans le but de faire la preuve de l'existence de terminologies françaises de qualité, aptes à exprimer les réalités industrielles et économiques modernes et dans un deuxième temps, identifier les lacunes existantes dans les domaines techniques afin d'enclencher une production terminologique motivée par la demande, qu'elle soit ponctuelle ou sectorielle, dans les domaines où il n'y a pas de réponse satisfaisante. Ainsi, l'Office de la langue française a conçu sa **Banque de terminologie** comme un système et un procédé de gestion des terminologies, tant sous l'aspect production que diffusion, dont on aura remarqué la complémentarité évidente.

Les inventaires et les productions terminologiques dans les secteurs traditionnels constituent ce qu'il convient d'appeler une opération de rattrapage. Le résultat de cette opération a permis de jeter les bases d'un système d'information terminologique qui contient approximativement aujourd'hui, d'une part 852 000 fiches portant sur environ 3 500 000 termes, témoins de la richesse du français et représentant les 37 secteurs d'activité économique du Québec, et d'autre part environ 41 000 références de répertoires terminologiques contenant du français.

La diffusion massive d'une banque de terminologie comme la **BTQ** a un «rôle important à jouer dans la normalisation sociale des terminologies» (Auger [1960]:30). Aussi avons-nous pensé à structurer sa diffusion dans une gamme de produits et de services en fonction des besoins du marché et en interaction avec la demande. Ainsi l'Office de la langue française offre-t-il à la population, grâce à sa **Banque de terminologie**, les produits et les services suivants :

1- l'accès direct par mode terminal;
2- l'accès en différé par des listages;
3- des consultations ponctuelles par téléphone;
4- des publications qui prennent la forme de répertoires bilingues ou de bibliographies;
5- et bientôt des minibanques personnalisées et constituées sur mesure par un téléchargement sélectif. Ces minibanques peuvent être exploitées en mode autonome sur mini ou micro-ordinateurs.

Les 3 500 000 termes que contient la mémoire de la **BTQ** concrétisent et harmonisent l'expression des rapports qui s'établissent dans la société, sur le plan social, industriel et économique.

«L'homogénéité des réponses que sa diffusion permet à travers un réseau bien développé et diversifié «par une gamme de produits et de services appropriés» aura en termes d'années un effet d'uniformisation très important sur les terminologies en usage au Québec» (Auger [1968]:30).

Et c'est principalement dans les communications institutionnalisées que l'apport de la **BTQ** aura été le plus tangible.

Par «communication institutionnalisée», il faut entendre, ainsi que le définit M. Jean-Claude Corbeil dans *L'aménagement linguistique du Québec* (Corbeil 1975:79) :

> «L'acte, le plus souvent anonyme ou impersonnel, par lequel une institution entre en relation avec des personnes, soit en tant que membres de cette institution, soit dans la relation employeur- employé, ou encore en qualité de clients, d'auditeurs ou de spectateurs».

Ainsi que le démontrent les rapports de gestion du réseau des abonnés de la **BTQ**, les maisons d'enseignement, l'administration publique et les entreprises représentent plus de 97% des clients du système en accès direct. Ils consomment 98% du temps d'interrogation et ils font 99% des transactions dans le système sous forme de commandes.

Cette répartition de la consommation de l'information terminologique a des incidences directes sur la qualité et la quantité des termes techniques utilisés dans la langue de l'enseignement des métiers et des professions, ainsi que sur celle des sciences et des techniques. Dans l'administration publique, l'information terminologique agit de la même façon sur la langue dans laquelle sont rédigés les lois, les règlements, les formulaires et la documentation à l'usage des citoyens. Enfin, dans les entreprises, l'utilisation de l'information terminologique permet d'assurer en français la gestion interne, d'exprimer et de réaliser des choix technologiques, d'assurer en français des relations avec la clientèle par l'étiquetage, les modes d'emploi des produits, les garanties, etc.

Est-il besoin de dire que les initiatives de l'Office de la langue française, en ce qui a trait à la **Banque de terminologie**, s'inscrivent dans la démarche de l'application d'une politique linguistique qui constitue «l'ensemble des choix conscients effectués dans les rapports entre langue et vie sociale et plus particulièrement entre langue et vie nationale» (Calvet 1967:155). Ces initiatives sont des exemples de la mise en oeuvre des moyens nécessaires à l'application d'une politique linguistique, moyens dont l'ensemble constitue ce qu'on appelle la planification linguistique qui s'exprime en deux volets, soit par des programmes d'action sur la langue (dont la terminologie est une des composantes essentielles), soit par des programmes d'action visant les rapports entre les langues.

Les enjeux technologiques et linguistiques

Ainsi que l'exprime M. Léon Dion, dans *L'État et la planification linguistique* : «Les langues, dans les sociétés industrielles avancées, en tant que véhicules de communication, continuent à être étroitement associées à la technologie ... en déterminant de façon plus directe que dans le passé les rôles sociaux les plus prestigieux.» (Dion 1981:22). Au Québec, l'émergence des nouvelles préoccupations sociales a été suivie d'un développement sans précédent de la technologie. Et pour exprimer le développement de la technologie, de nouveaux termes doivent être créés quotidiennement.

La **Banque de terminologie du Québec** doit donc évoluer au même rythme que les activités liées au marché des 37 secteurs d'activité économique du Québec. Le déploiement de nouvelles technologies, en particulier celles qui sont liées aux biotechnologies, aux techniques informatiques et à la microélectronique, fait qu'il existe peu de métiers et de professions qui ne soient présentement l'objet d'un questionnement d'ordre technologique, affectant dans l'immédiat la terminologie en usage. Un bon nombre de ces professions ont une durée de vie bien comptée, d'autres sont en pleine transformation ou en pleine émergence, notamment dans le secteur d'activités qui nous touche de près, nous, terminologues, traducteurs, celui des industries de la langue qui, incidemment, évolue sous l'impulsion d'ingénieurs-cogniticiens. A l'heure actuelle, des pans entiers du savoir sont en train de se renouveler. Qu'on pense seulement à l'électronique qui, malgré ses origines diverses, était dotée au début du siècle d'un vocabulaire bien français jusqu'à la révolution microélectronique dont témoigne l'invention du transistor, en 1948, dans les laboratoires Bell. Par la suite, ce fut la course effrénée : invention du circuit intégré en 1958 par **Fairchild et Texas Instrument**, le microprocesseur en 1971 par **Intel**, et voilà qu'après les tubes et les semi-conducteurs l'**American**

Physical Society annonce la possibilité de fabriquer des supra-conducteurs qui fonctionnent à la température ambiante (Pharand 1987:5). C'est la troisième révolution de la microélectronique qui se profile à l'horizon.

A l'heure actuelle, la microélectronique est conçue et libellée en anglais. Phénomène sans conséquence, pensez-vous! Après tout, elles ne sont pas si nombreuses les entreprises qui fabriquent des composantes électroniques au Québec. Mais quelles sont les entreprises aujourd'hui qui ne font pas appel à la microélectronique à une étape ou à une autre du processus de production? Selon le président du Conseil de l'industrie électronique du Québec, M. Pierre Pharand, «toutes les entreprises, même les toutes petites, y font appel et l'expérience nous montre que, dans ce domaine, la langue utilisée est d'abord l'anglais» (Pharand 1987:5).

L'évolution de la microélectronique est une véritable épopée qui propulse l'évolution scientifique dans le domaine de l'aérospatial et a des retombées techniques dans la vie de tous les jours par l'utilisation d'une panoplie de dispositifs électroniques.

L'exemple que nous venons de donner n'est pas unique; il pourrait être le modèle descriptif de ce qui arrive avec les biotechnologies et les techniques informatiques qui se répercutent dans la plupart des secteurs d'activités, que ce soit dans l'industrie automobile, l'imprimerie, l'édition ou encore dans l'un ou l'autre domaine du secteur tertiaire, par exemple dans les services financiers et bancaires.

Cela affecte considérablement le volume des terminologies à mettre à jour et leur durée de vie. Au cours de l'année 1986-1987, les terminologues de l'Office ont fait 152 000 mises à jour dans le système **BTQ**, soit par l'ajout de 36 200 nouvelles fiches, soit par la modification ou par la suppression de 115 700 fiches. La mise à jour a donc touché 18% du contenu de la banque.

Les horizons du développement de la BTQ

En terminologie, les problèmes sont rarement réglés une fois pour toutes. Il faut vérifier périodiquement les équivalences des termes, vérifier et adapter leur définition, vérifier les séries synonymiques, etc. Cela témoigne dans une certaine mesure de la santé et de la vitalité de la langue.

Si nous nous tournons vers l'avenir et que nous nous demandons maintenant comment nous pourrions parcourir le chemin qu'il reste à faire, il est clair que ce que nous venons de dire conditionne la nature et la portée de nos priorités en matière de gestion de l'information terminologique pour les prochaines années.

Il faut de toute urgence et de façon concertée avec nos partenaires (les entreprises, l'administration publique, le réseau de l'enseignement) :

1 - Orienter le système **BTQ** de production terminologique vers les sciences et les technologies porteuses de l'innovation industrielle, et au premier chef les biotechnologies, les techniques informatiques, la microélectronique, et engager une production massive.
2 - S'appliquer à renouveler les vocabulaires traditionnels en fonction du leadership qu'exercent les nouvelles technologies sur les métiers et les professions, à la fois dans les secteurs industriels et dans les services.

Dès lors, une question se pose : comment peut-on s'organiser efficacement pour que la **Banque de terminologie du Québec** constitue une valeur sûre dans dix ans? Comment doit-on s'organiser pour suivre l'évolution de 3 000 000 à 4 000 000 de termes tout en ajoutant quelques dizaines de milliers de termes par année pour dénommer les inventions nouvelles?

Le temps manque aujourd'hui pour couvrir de façon exhaustive tous les aspects de la question, et surtout le volet des techniques informatiques.

Je voudrais tout simplement dire et répéter qu'à mon avis la solution se trouve dans un traitement intelligent de la langue naturelle en empruntant à l'intelligence artificielle et aux systèmes-experts la combinatoire requise pour mettre au point un système d'aide à la recherche terminologique, et principalement pour le dépouillement automatique des textes techniques. Cela constitue notre troisième priorité. Elle est rendue nécessaire pour réaliser efficacement les deux premières.

Le Québec bénéficie d'une bonne expertise en la matière et les bases du système existent. Comme l'exprime Richard Parent, du Bureau central de l'informatique au gouvernement du Québec, il s'agit de poursuivre «la recherche en l'actualisant dans les projets-pilotes (à des fins terminologiques) ... La partie système-expert a été développée ... une batterie de structures de données, divers modèles d'exploration, des séries d'automates, d'opérations, de fonctions, le sont également ... , (nous avons de plus) l'analyseur lexico-syntaxique du français qui contient beaucoup de données et de connaissances sur la langue française ... De nouveaux efforts doivent donc être consentis, notamment pour la construction d'un module de traitement linguistique ...» (Parent:15) A court terme, une collaboration pourrait être envisagée avec l'Université, le Bureau central de l'informatique et l'Office et il est permis d'envisager des résultats concrets dès 1990.

Le Québec a acquis une nouvelle maturité, une nouvelle sécurité, une nouvelle confiance dans son développement linguistique, de sorte que «l'Office de langue française peut se préoccuper surtout (à l'avenir) de prévention et de promotion, de concert avec ses partenaires» (Laporte 1987:19). C'est ainsi que s'exprimait le président de l'Office, M. Laporte, dans la conférence qu'il prononçait dans ce même Carrefour.

C'est également en pensant à cette façon d'entrevoir l'avenir que l'on peut affirmer que le Québec a besoin d'un système de production et de diffusion terminologiques à la fine pointe de l'actualité, défi que l'Office et sa **Banque de terminologie** devront relever pour assurer le rayonnement de la langue française dans toutes les sphères d'activité économique du Québec, aux horizons de l'an 2000.

BIBLIOGRAPHIE

ALLAIRE (Yvon) et (Roger) MILLER. *L'entreprise canadienne et la loi sur la francisation du milieu de travail*, Montréal, Institut de recherches C.D. Howes, 1980, 80 p.

AUGER (Pierre). *La terminologie au Québec*, quinze ans d'aménagement linguistique, [1968], 38 p.

CALVET (Jean-Louis). *La guerre des langues et les politiques linguistiques*, Paris, Payot, 1987, 294 p. (Langages et sociétés)

CORBEIL (Jean-Claude). *L'aménagement linguistique du Québec : perspective historique de la question : description des options linguistiques qui sous-tendent l'action de l'Office de la langue française du Québec*, Québec, Éditeur officiel du Québec, c1975, 57 p. (Études, recherches et documentation).

DION (Léon). *«L'État, la planification linguistique et le développement national»*, in *L'État et la planification linguistique*, s.l., Office de la langue française, 1981, (Langues et sociétés), t.1, p. [13] - 35.

LAPORTE (Pierre-Étienne). *La francisation des entreprises : une innovation québécoise*, 1987, 21 p.

MARTIN (André). *L'État et la planification linguistique.*

PARENT (Richard). *Un module linguistique pour l'informatique gouvernementale des années 1990*, communication présentée lors de la Deuxième Conférence des chefs d'État et de Gouvernement des pays ayant en commun l'usage du français sous le thème : Industries de la langue, 5 septembre 1987, p. 15.

PHARAND (Pierre). «Préface» in CAMPBELL, Carmen, et NOLET, Andrée, *Lexique de la microélectronique*, [Montréal], Office de la langue française, 1987, (cahiers de l'Office de la langue française. Terminologie technique et industrielle), p. [5] - 6.

TERMIUM : OUTIL MODERNE DE COMMUNICATION[1]

Winston G. Wells
Bureau des traductions
Secrétariat d'État

Dans un pays bilingue et multiculturel comme le Canada, il ne faut pas se surprendre de voir un nombre considérable de gens s'intéresser aux industries de la langue. Il s'agit d'un secteur dans lequel les Canadiens possèdent des compétences reconnues et ont accompli de nombreuses réalisations. Les deux grands groupes linguistiques du pays se sont donné des outils efficaces de communication. Le Canada n'aurait sans doute pas pu atteindre le degré de cohésion que nous observons aujourd'hui sans la présence, constante et dynamique, de milliers de personnes engagées dans la promotion de la communication dans les deux langues officielles du pays.

Rôle du Secrétariat d'État en matière linguistique

Le Secrétariat d'État du Canada joue un rôle important dans le processus de transmission de la communication entre les communautés anglophone et francophone du pays. Un des grands mandats qui ont été confiés à ce ministère consiste précisément à favoriser, par divers programmes et activités, l'épanouissement, dans sa langue, de chacun des deux groupes linguistiques. Ce mandat découle de l'adoption, en 1969, de la *Loi sur les langues officielles* qui garantit à tout citoyen canadien l'accès aux services gouvernementaux dans la langue officielle de son choix. De plus, tous les documents qui sont rendus publics par l'administration fédérale doivent être publiés en français et en anglais. De telles dispositions de la Loi obligent l'administration à se doter de services de soutien linguistique importants. Le secteur **Langues officielles et traduction** du Secrétariat d'État constitue le centre névralgique sur lequel s'appuie le gouvernement canadien pour mettre en oeuvre sa politique linguistique.

Le Bureau des traductions

Créé par une loi du Parlement en 1934, le Bureau des traductions du Secrétariat d'État dessert l'ensemble de l'appareil administratif fédéral. Il fournit des services de traduction, d'interprétation et de terminologie, ainsi que des services linguistiques aux ministres, députés, sénateurs et aux quelques 260 000 fonctionnaires de tous les ministères et organismes qui en dépendent. Si les services de traduction et d'interprétation ne sont destinés qu'à la fonction publique, il en va autrement pour la recherche terminologique et les conseils linguistiques qui rejoignent quant à eux l'entreprise privée et le grand public. L'ampleur de la tâche à accomplir a incité le gouvernement canadien à mettre en place une infrastructure capable de répondre à la demande croissante de services linguistiques, en plus de se doter d'outils à la fine pointe du progrès.

---

[1]Communication présentée lors de la Deuxième Conférence des chefs d'État et de Gouvernement de pays ayant en commun l'usage du français, sous le thème : Industries de la langue, Québec, 5 septembre 1987.

La banque de données linguistiques Termium

Parmi les outils modernes destinés à faciliter la communication au sein d'une société bilingue et multiculturelle, il y a la Banque de données linguistiques du gouvernement canadien, connue sous le nom de **Termium**. Mise au point par une équipe de chercheurs de l'université de Montréal au début des années 1970, puis prise en charge par le Secrétariat d'État en 1975, la banque **Termium** renferme aujourd'hui près de 900 000 fiches. Maintenant dotée d'un logiciel de troisième génération, elle a subi de nombreuses transformations depuis son acquisition par le Bureau des traductions. Aux 150 000 fiches de départ sont venues s'ajouter des centaines de milliers d'autres fiches recueillies auprès de l'ensemble des traducteurs de l'administration fédérale et de plusieurs universités du pays ainsi qu'au sein des services linguistiques de grandes entreprises canadiennes. Depuis plus de dix ans, une importante équipe de terminologues s'affaire à mettre de l'ordre dans toutes ces données, en plus de verser régulièrement dans **Termium** les nouveautés terminologiques rencontrées au gré de travaux de recherches ponctuelles et thématiques.

Destinée d'abord à répondre aux besoins d'un nombre considérable de traducteurs au service de l'État, qu'ils soient fonctionnaires, pigistes ou entrepreneurs, la banque **Termium** est composée de quatre grandes bases de données particulièrement bien adaptées aux besoins de cette clientèle. Il s'agit de la base *Appellations* qui renferme plus de 75 000 titres; de la base *Multilingue* qui contient plus de 33 000 fiches et où l'allemand et l'espagnol côtoient le français et l'anglais; de la base *Terminologie*, riche de plus de 775 000 fiches rédigées en anglais et en français et de la base *Difficultés de traduction*, offrant environ 10 000 solutions à autant de problèmes que peut rencontrer le traducteur. A peu près tous les domaines de la connaissance sont couverts, dans des proportions variées, par cette masse colossale de données linguistiques. Cette dernière caractéristique est sans doute celle qui a le plus contribué à étendre la diffusion de **Termium** auprès d'individus, d'entreprises et d'organisations voués à la promotion de la communication tant en français qu'en anglais. Le réseau **Termium**, qui compte présentement 288 terminaux, dont 27 à l'étranger, pour un total d'environ 3 000 utilisateurs, est loin d'avoir atteint son potentiel maximal de diffusion. Des contraintes financières et techniques nous ont empêché jusqu'à maintenant d'étendre davantage son rayonnement. Cette situation est cependant appelée à changer grâce, principalement, aux progrès technologiques et à l'immense potentiel de **Termium**.

Situation privilégiée du Canada

Le Canada, en tant que pays bilingue, a le bonheur d'appartenir à deux grandes communautés linguistiques internationales, celles de la langue anglaise et de la langue française. En fait, il y a peu de pays bilingues qui possèdent cet atout. Grandes langues de communication, le français et l'anglais possèdent un statut des plus enviables. Cette particularité du tissu linguistique canadien a permis au Canada, depuis quelques décennies surtout, de jouer un rôle de premier plan dans des domaines tels que la traduction automatique et la constitution de banques de terminologie informatisées. Ce n'est donc pas un hasard si le Canada possède le seul système de traduction complètement automatique de haute qualité, le système **MÉTÉO**, mis au point par l'équipe **TAUM** de l'université de Montréal, ainsi que deux des cinq grandes banques de terminologie qui existent dans le monde occidental, soit la **Banque de terminologie du Québec** et **Termium**. Ces outils modernes de communication, mis au service de deux grandes langues largement répandues sur la planète, ont permis au Canada d'établir de multiples relations avec une foule d'organismes internationaux et de mettre ainsi en évidence son savoir-faire dans le domaine des technologies de pointe appliquées aux industries de la langue.

Termium au service de la promotion des langues officielles

Devenu, au fil des années, un outil de communication et d'échange international, **Termium** n'a pas pour autant atteint son plein épanouissement national. Le rôle qui lui est dévolu, celui de pro-

mouvoir le bon usage des langues officielles au Canada, reste encore à être développé. Certaines contraintes, principalement financières, ont obligé le Bureau à freiner l'expansion du réseau **Termium**. Cette situation pourrait bien changer du tout au tout grâce aux progrès récents de la technologie, principalement en ce qui concerne la mise au point du disque optique numérique ou **CD-ROM**. Les gros systèmes informatiques ont des limites quant au nombre d'utilisateurs qu'ils peuvent desservir simultanément. En plus d'être coûteux, autant à l'achat qu'à l'entretien, ces systèmes ont parfois du mal à s'adapter à la dispersion géographique de leurs utilisateurs.

La présence de six fuseaux horaires au Canada est un des facteurs qui militent en faveur de l'accès à **Termium** par d'autres moyens que l'interrogation en direct de l'ordinateur, qui se trouve dans la région d'Ottawa-Hull. D'autres facteurs plus importants, tels que la limite de puissance du système et le nombre restreint d'utilisateurs capables d'interroger simultanément **Termium**, incitent le gouvernement canadien à explorer de nouvelles avenues afin de rendre disponible à un plus grand nombre de citoyens cette aide précieuse à la rédaction dans les deux langues officielles du Canada.

La technologie du disque optique, pour ne citer que celle-là, a devant elle un avenir prometteur. De nombreux secteurs d'activité s'y intéressent. Le stockage des données terminologiques semble être un de ces domaines où l'application de la technologie du disque compact est possible. Un projet-pilote a été mis sur pied par le Secrétariat d'État et consiste à offrir à un nombre limité d'utilisateurs du Bureau des traductions et du secteur privé un disque optique sur lequel se trouve tout le contenu de la banque **Termium**. Il suffit de placer le disque dans un lecteur de disque optique relié à un micro-ordinateur pour avoir accès sur le champ aux quelque 900 000 fiches de **Termium**, soit environ deux millions de termes. Les tests, qui s'étendront sur plusieurs mois, vont nous permettre de savoir dans quelle mesure cette nouvelle technologie pourra servir à augmenter le nombre d'abonnés à **Termium**. Il y a tout lieu d'être optimiste quant aux résultats de l'expérience en cours.

Il a été mentionné, au début de ce texte, le nombre de 288 terminaux qui constituent le réseau actuel de **Termium**, dont 261 sont installés au Canada, la majorité de ceux-ci se trouvant au Bureau des traductions du gouvernement canadien. Les traducteurs du Secrétariat d'État, tout comme les terminologues, ont établi une communication constante avec **Termium**. La communication entre les deux langues officielles dans la fonction publique canadienne passe souvent par **Termium** et ce, par traducteurs et terminologues interposés. Comment ne pas envisager, grâce au disque optique, une plus grande diffusion de **Termium** au sein même de la fonction publique, auprès des centaines de rédacteurs, correcteurs et autres fonctionnaires appelés à travailler régulièrement dans les deux langues officielles de l'administration fédérale? Comment ne pas songer également à un autre marché potentiel considérable et en pleine expansion, celui des classes d'immersion de langue française qui jouissent d'une grande popularité un peu partout au Canada? A l'heure de l'informatique, élèves, étudiants et enseignants auraient tôt fait d'exploiter toutes les subtilités du logiciel **Termium**. Il ne faut pas oublier les centaines d'entreprises des secteurs des services et de l'industrie qui sauraient trouver en **Termium** un outil moderne et efficace, propice à faciliter la communication tant en anglais qu'en français. Pourquoi le grand public n'aurait-il pas accès éventuellement à **Termium** comme il peut déjà avoir accès à certains services électroniques? **Termium** sur un **Minitel** canadien, pourquoi cela ne serait-il pas possible? C'est, à mon avis, dans cette optique qu'il faut envisager le rôle que doit jouer **Termium** dans la promotion des langues officielles au Canada.

Termium, outil de communication international

Le réseau international de **Termium** compte actuellement 27 points d'accès, situés essentiellement aux États-Unis et en Europe occidentale. D'importants centres de services linguistiques américains et européens ont vu l'intérêt qu'ils avaient à se relier à **Termium** à cause du contenu varié de ses bases de données, de la qualité et de la quantité de ces données. **Termium** est devenue une monnaie d'échange qui permet d'obtenir des abonnés à l'étranger de précieuses données servant à mettre à jour les fichiers. De plus, **Termium** contribue à mettre en place tout un réseau de collaborateurs et

d'interlocuteurs capables de comprendre, et même de défendre, les positions canadiennes dans les grands forums internationaux s'occupant des industries de la langue, principalement en ce qui a trait à la recherche terminologique.

Plusieurs rencontres avec les représentants des autres grandes banques de terminologie ont permis d'établir certains mécanismes propices aux échanges de renseignements sur les travaux en cours dans les grands centres de terminologie, sur les difficultés rencontrées de part et d'autre et sur les moyens de collaborer plus étroitement à l'établissement d'un véritable réseau international de banques de données terminologiques.

Conclusion

Un des grands avantages du développement de l'informatique est précisément de faciliter le dialogue entre les peuples des divers continents. La rapidité avec laquelle se font maintenant les communications permet d'envisager l'établissement de nombreux contacts partout à travers le monde. **Termium**, en tant qu'outil destiné à faciliter le dialogue entre les langues, entend participer pleinement à l'essor technologique en cours afin d'offrir, au plus grand nombre possible de clients, tant au Canada qu'à l'étranger, les trésors linguistiques qu'elle renferme. Déjà implanté aux États-Unis et dans plusieurs pays d'Europe, le réseau international de **Termium** ne saurait être complet sans la présence, en son sein, d'un certain nombre de pays africains. Déjà, une première terminologique a été réalisée dans le cadre du deuxième sommet francophone par la publication de deux vocabulaires rédigés en ikinyarwanda, en français et en anglais. Ces deux ouvrages portent sur l'industrie agro-alimentaire en Afrique. Grâce à la collaboration de l'Office de la langue française du Québec et au soutien technique et financier du Secrétariat d'État du Canada, le professeur rwandais Léon Mugesera, par l'entremise du Réseau international de néologie et de terminologie, peut maintenant offrir à ses compatriotes ainsi qu'à l'ensemble de la francophonie, le fruit de ses travaux terminologiques menés à l'université Laval, ici-même à Québec. Par ailleurs, il convient de signaler que le lancement de ces deux publications se fera aujourd'hui même, à 18h30, à l'amphithéâtre du manège militaire. Il s'agit là du résultat d'une heureuse collaboration Nord-Sud, collaboration que nous nous devons tous de promouvoir au plus haut point.

Il y a sans doute bien des façons de faire progresser la communication et les échanges au sein de la francophonie internationale. Le Secrétariat d'État du Canada, à l'exemple du rôle qu'il joue dans notre beau et grand pays, entend favoriser le rapprochement des peuples par des moyens tels que la diffusion de **Termium** et sa participation au Réseau international de néologie et de terminologie. Mis au service de la communication, ces outils, si modestes soient-ils, n'en contribueront pas moins, nous l'espérons, à rendre plus harmonieux le grand dialogue des nations.

REMARQUES SUR L'AMÉNAGEMENT DU STATUT DU FRANÇAIS EN INFORMATIQUE[1]

Jean-Claude Boulanger
Langues et linguistique
Université Laval

Au détour des années 1950 et dans le sillage de la guerre finissante, on commença à percevoir l'accélération phénoménale du développement de la technologie de l'informatique et son insertion dans les ordres spécialisés du savoir. Quarante ans plus tard et à la veille de la conclusion du deuxième millénaire de l'humanité, un quatuor de termes reliés au même technolecte semble dominer l'univers des connaissances et témoigne de l'impact de l'informatique sur les sociétés. Il s'agit des termes *ordinateur* (1955) (Marcellesi, 1979:176-180) et de son rejeton *micro-ordinateur* (1971), *informatique* (1962) et de son satellite immédiat *micro-informatique* (vers 1980), *logiciel* (vers 1970) et de l'adjectif *artificiel* que l'on trouve associé à quelques substantifs (ex. *intelligence*, *langage*). A sa manière, chacun a dominé sa décennie et il est indéniable qu'ils font désormais partie intégrante de l'histoire du français; à double titre d'ailleurs : premièrement parce qu'ils sont devenus des termes techniques indispensables, deuxièmement parce qu'ils se sont coulés dans le moule de l'usage général.

Bien entendu, il est d'abord fait référence ici à la macro-informatique souvent dénommée *informatique professionnelle*. Les mutations technologiques successives ont pris une ampleur nouvelle au commencement des années 1980 avec l'avènement de la micro-informatique qui a rendu matériels et logiciels disponibles et accessibles pour toutes les catégories d'usagers, spécialistes ou profanes. La micro-informatisation entraîna à sa suite une amorce de démystification de la technologie et le début de la banalisation de la terminologie propre au secteur. Une bonne partie du vocabulaire de l'informatique n'est plus l'apanage exclusif des informaticiens, des penseurs, des concepteurs, des publicitaires, des diffuseurs et des vendeurs de produits. L'aggiornamento informatique trouve son achèvement dans les dictionnaires de langue pour adultes ou pour élèves. Ainsi la nouvelle édition du *Micro-Robert* qui sera mise en marché très bientôt, répertorie tout le vocabulaire de l'informatique passé dans l'usage commun. On en fait même un argument publicitaire. La visibilité de l'informatique dans les foyers oblige les responsables à se pencher sur la «déjargonisation» du lexique. En particulier, il faudra autopsier le statut de l'anglais comme langue véhiculaire de la technologie micro- ou macro-informatique.

Les quelques problèmes qui paraissent jouer un rôle majeur dans les rapports entre les concepts d'«informatique» et de «francophonie» seront abordés ci-après. La position adoptée ici est celle du linguiste, qui use à l'occasion des outils informatiques, surtout de la micro, et qui a des préoccupations en ce qui concerne les capacités d'une langue de faire face au renouvellement de son fonds par le recours aux formes autochtones inédites ou existantes.

A travers les siècles, le français a toujours su nationaliser les terminologies qui le nécessitaient. Pourquoi en serait-il autrement aujourd'hui alors qu'il paraît mieux équipé que jamais pour le faire? Après la période de mise en place de l'arsenal des mécanismes de formation des mots en français au Moyen Age, toutes les époques se sont heurtées au besoin de contrer les langues étrangères (arabe, latin, italien, anglais) ou de combattre des jargons scientifiques, juridiques, scolastiques, ecclésiastiques et théologiques, et même techniques. Chaque étape de l'histoire du français atteste

[1]Communication présentée lors du 56e Congrès de l'ACFAS, Moncton, 9-13 mai 1988.

amplement des efforts d'émancipation réussis : au XVI[e] siècle, la langue du peuple commence à prendre le pas sur le latin dans le domaine juridique (Édit de Villers-Cotterêts en 1539), en médecine (chirurgie), en astronomie, en musique, ainsi de suite; au XVIII[e] siècle, les sciences naturelles et la chimie sont mieux perçues en français. Peu importe les époques, un seul motif explique l'emploi du vernaculaire : l'usager ordinaire de la langue générale et le consommateur de technolectes ne peut plus décoder les terminologies savantes ou étrangères. Ainsi donc des pans entiers du lexique spécialisé ont échappé peu à peu aux langues doctes en raison de l'incapacité des utilisateurs de parler ou de comprendre ces langues (Boulanger 1988b:4-9). La force du besoin a servi de catalyseur afin d'encourager la naturalisation lexicale.

Ce bref tracé historique sert de trame au problème du langage informatique d'aujourd'hui, qui se ramène en fait à savoir si le français peut exprimer sans contrainte les concepts de ce vaste domaine ou s'il doit être à la remorque de l'américain, avant d'être phagocyté par lui. La réponse à cette interrogation fondamentale indiquera si l'édifice du français risque de s'effondrer, à tout le moins de se fissurer, ou si l'alarme n'est rien d'autre qu'une suite de fantaisies de linguistes.

Quelques facteurs seront examinés, dont les uns jouent en faveur du maintien et du développement du français et dont les autres, sans être négatifs, ont néanmoins des incidences non négligeables sur ses structures. Parmi ces facteurs, un seul est d'origine purement linguistique. Il ne sera abordé que sommairement en raison du fait que la langue est constamment présente en filigrane dans les autres éléments. De fait, elle les cimente tous en un réseau de liens très complexes.

2. L'informatisation

L'informatique est un secteur de connaissance très vaste et intimidant pour le néophyte ou le non-spécialiste. C'est aussi un domaine mythifié par la langue et le vocabulaire employés, à savoir l'anglais ou plus précisément une certaine forme d'anglais qui s'apparente à une langue quasi artificielle ou à un métalangage. Au départ, l'informatique est une technologie importée dans le monde francophone. Une chercheuse constatait en 1973 que 90% du marché français était occupé par des constructeurs américains (Marcellesi 1973:59). Le vocabulaire a naturellement suivi la chose de sorte que «l'emploi, qui semble abusif au profane, de termes anglo-américains, la difficulté même de la technique et le mystère dont elle est souvent entourée aux yeux du non-initié, rendent ce langage particulièrement incompréhensible pour qui n'a aucune notion d'anglais ou d'informatique» (Marcellesi 1973:59). Malgré ses ouvertures franches, l'informatique demeure une chasse gardée pour nombre d'utilisateurs potentiels. Les publicitaires eux-mêmes ne s'emploient guère à résoudre le mystère. Grande diffuseuse de mots, la publicité «exploite et démultiplie toutes les dénominations susceptibles de flatter les mythes ou les préjugés des clients virtuels. Or elle considère [...] les termes étrangers comme des «mots-qui-font-vendre»» (Quemada 1978:1228).

Depuis longtemps et dans plusieurs milieux, on souhaite que des efforts sérieux soient accomplis pour rendre l'informatique plus accessible, plus compréhensible au vaste public. Derrière ce désir relié à la diffusion du savoir technologique, se profile le seul et unique moyen rentable de réaliser la tâche. Il consiste à offrir la terminologie en français ou plutôt à rendre fonctionnel le vocabulaire disponible. En réalité, c'est plutôt la micro-informatique qui est pointée ici puisqu'il appert que la macro-informatique demeurera sans doute encore longtemps l'apanage des spécialiste informaticiens. Si la macro-informatique représente une micro-culture technicienne, à l'inverse la vulgarisation plus rapide de la micro-informatique façonne une macro-culture nouvelle.

3. La miniaturisation

Les générations d'ordinateurs s'enchaînent avec comme double objectif «la nécessaire décroissance des dimensions des machines et la recherche de marchés nouveaux» (Goujon 1985:400), sans compter le désir d'augmenter la puissance des logiciels. Afin de répondre à ce voeu et depuis le

début de la décennie, la cartographie informatique se ramifie en de multiples sous-secteurs et applications qui favorisent sa diffusion sociale. Ainsi en va-t-il de la domotique, de la servitique, de l'éditique, de l'autotique, alors que des domaines de l'activité traditionnelle se micro-informatisent. Le mouvement vers la miniaturisation des équipements s'amorce vers 1960. Le premier micro-ordinateur du monde, le Micral, apparaît en 1972 tandis que la première boutique de micros ouvre en 1975 à Los Angeles (Goujon 1985:401). A l'heure actuelle, les incidences de ce phénomène sur les avancées ou le recul de la francisation sont difficilement mesurables. Cependant, les attentes sont nombreuses comme le démontrent les recherches et les enquêtes menées auprès des usagers de l'infographie dans les domaines artistiques (arts visuels en particulier) et culturels (Turcotte 1988).

Le développement fulgurant et récent de quelques technologies a entraîné le passage obligé de la macro-informatique à la micro-informatique. La miniaturisation a donné une nouvelle impulsion aux recherches centrées sur l'intelligence artificielle, sur la **TAO**, sur la **TERAO** et sur la linguistique, pour mentionner quelques domaines qui nous concernent de plus près. A ce chapitre, il est clair que les «résultats de l'informatique linguistique favorisent le développement, la fabrication et la commercialisation des technologies et des produits tout à fait révolutionnaires, qui rendent d'indéniables services dans diverses sphères d'activité» (Boulanger 1988b:19-20). Simultanément, le poids et l'influence de la langue anglaise se sont accrus dans le monde. L'anglais consolide sa place privilégiée et sa position de force par rapport aux autres langues. Pressées de s'équiper et de fonctionner avec ces nouveaux outils révolutionnaires, peu de sociétés se sont penchées sur l'impact que cette technologie provoque sur l'ensemble de la ou des langues en usage dans ces communautés ou sur leur territoire. Éblouis par les possibilités nouvelles et par la doxa informatique, quelques pays africains francophones ont introduit l'anglais dans leur projet d'aménagement linguistique dont le premier objectif consistait à promouvoir les langues nationales. Des projets prennent du retard ou sont repoussés, ils passent au second plan pour se dissoudre enfin dans le faux brouillard de la modernisation. L'industrialisation et la planétarisation massive de l'anglais creusent davantage le fossé entre les langues. L'écart s'agrandit en faveur de la langue des producteurs de systèmes et d'outils. L'informatique grande ou petite, matérielle ou logicielle, est pensée, produite et surtout commercialisée par des entreprises américaines. Les objets sont exportés vers d'autres milieux consommateurs qui les acceptent tels quels, avec dans leur sillage les termes allogènes demeurés intacts (les emprunts) ou légèrement retouchés pour prendre un visage en apparence français (les calques) (Boulanger 1988a). A l'occasion, quelques notions parviennent à donner naissance à des créations internes originales qui sont à leur tour réexportées avec succès vers l'anglais (ex. *télématique* -> *telematics*), démontrant qu'un aller-retour est parfois possible.

Mais il reste qu'on dit et écrit fréquemment que l'informatique est une «langue» peuplée d'emprunts et de calques, une «langue» difficile, voire impossible à traduire. Quand on observe l'augmentation de l'amplitude de l'informatique dans les activités quotidiennes, quand on constate l'accroissement de la vitesse de pénétration des trouvailles technologiques dans la vie courante d'une année sur l'autre, on est en droit de se poser des questions sur l'avenir du français dans l'espace micro-informatique. D'autant que jusqu'à récemment, l'informatique s'était limitée à jouer le plus souvent un rôle de support pour transmettre l'information lexicale. Désormais, elle devient en plus l'information lexicale en soi, justifiant aux yeux de ses ardents défenseurs qu'elle pourrait aspirer au statut de langue. Nombreux sont ceux qui croient que l'empire informatique s'identifie à une «langue» multinationale qui dépasse le strict niveau lexical pour fonder sa propre syntaxe. Le lexique, y compris la variation d'entreprises, la grammaire et la syntaxe se combinent pour générer un pur métalangage, un code aux règles internes bien contrôlées et appartenant aux seuls tenants de ce savoir. Le technolecte informatique ne risque-t-il pas ainsi de se transformer en un artifice qui deviendra bientôt si sophistiqué et marginalisé que nul ne le comprendra si ce ne sont les purs ingénieurs informaticiens?

4. La banalisation

La miniaturisation ou l'atomisation a eu comme effet secondaire de permettre à l'informatique de percer le marché domestique et d'élargir sa palette d'influence. La micro-informatique constitue

désormais un facteur culturel et social non négligeable, qui a des répercussions sur de multiples gestes quotidiens des individus non préparés à affronter cette technologie. Les logiciels se métamorphosent en produits de consommation banals, disponibles partout, concurrentiels, faciles d'utilisation et qui s'adaptent aux besoins des utilisateurs. De plus, il sont de moins en moins coûteux. Tout propriétaire de micro devrait «pouvoir utiliser l'ordinateur sans rien connaître du principe du traitement électronique de l'information ou de la programmation» (Meney et coll. 1987:2). D'où le succès incontestable des mécanismes comme la souris et les menus didactiques. Avec un minimum de débrouillardise, les «utilisateurs domestiques, ceux qui n'ont pas encore renoncé, acquièrent peu à peu une culture informatique suffisante qui leur permet de maîtriser peu ou prou le fonctionnement de leur machine» (Goujon 1985:403).

La diffusion sociale de l'informatique entretient semble-t-il un écart entre les usagers ordinaires et les professionnels de la technologie. Des recherches et des enquêtes récentes ont démontré que «parmi les catégories d'utilisateurs, seuls les professionnels et les informaticiens semblent préférer travailler avec des logiciels de langue anglaise» (*Rapport d'étape du comité Canada-Québec sur le développement du logiciel d'expression française* 1986:2). En d'autres mots, le simple citoyen se plaint ou se plaindra du rôle prédominant de l'anglais. Ceci pourrait avoir des conséquences économiques désastreuses lorsque le plafonnement des produits sera atteint et qu'il faudra trouver de nouveaux débouchés. La maîtrise du langage étant essentielle à la maîtrise des produits, il est évident que les rapports du langage avec la connaissance et la pensée, et avec la réalité et la pratique, doivent être soignés.

L'informatique, ou plutôt la micro-informatique, doit se franciser pour la bonne et unique raison qu'elle se banalise. Elle sort du cercle étroit et du despotisme des spécialistes. En se vulgarisant, elle engendre des exigences et des attentes en vertu des usagers qui se multiplient et qui l'utilisent comme un moyen, un outil et non comme une fin. En se transformant en objet quotidien, en étendant ses zones d'intervention, l'informatique perd son caractère d'alchimie, ce qui en fait une bonne candidate à la francisation.

5. La conscientisation

La terminologie est l'une des ressources linguistiques disponibles dont se servent les sociétés pour assurer leurs métamorphoses et leur avenir. «En vertu d'une logique interne, toute langue évolue et s'adapte au nouvel ordre de vie d'une société» (Boulanger 1988b:3). Mais quand il s'agit des langues de spécialité, il faut y adjoindre une condition : c'est la volonté collective d'agir sur la langue. Inscrire des termes, des néologismes dans le cycle de l'usage relève de considérations qui n'ont souvent pas de grandes résonances d'origine linguistique mais plutôt des sources d'origine sociale, dans le sens le plus englobant du mot.

L'organisation rationnelle et fonctionnelle des travaux repose toujours sur le principe volontariste et valorisateur. L'aménagement linguistique du vocabulaire de l'informatique passe par le développement et le «confortement» d'une idéologie en ce qui a trait à la terminologie et à la néologie françaises. Le terme *aménagement linguistique* désigne ici «l'ensemble des dispositions sociales qui influenceront le comportement linguistique des individus au sein de la société» (Corbeil 1978:157). Tandis que la néologie sera perçue comme le processus auquel recourent les individus pour créer des mots ou des sens nouveaux dans une langue. Les innovations linguistiques servent à plusieurs fins parmi lesquelles on peut retenir les deux plus visibles : 1 - augmenter naturellement le stock lexical d'une langue et 2 - permettre de transiter d'une langue à une autre en vertu d'un principe comme celui de la francisation. Les technolectes s'alimentent des dénominations des deux types. L'informatique paraît construire sa terminologie en s'appuyant davantage sur le second modèle, c'est-à-dire le transfert par emprunt ou par calque, entretenant ainsi l'image du palimpseste de l'anglais (Boulanger 1988a et Humbley 1987).

Lorsque le phénomène de la conscientisation est évoqué, le premier objectif à atteindre est certainement l'infléchissement des comportements fixistes et élitistes du corps informatique. C'est le seul moyen de faire face à la francisation de la pensée et du vocabulaire de ce secteur technolectal et ainsi de cesser de vouloir concurrencer l'anglais qui s'arroge le monopole de cette terminologie en ce moment. Ne vaudrait-il pas mieux devenir agressif et offensif en français plutôt que de se défendre contre l'envahisseur en même temps que l'on défend l'envahisseur? Superposer le français à l'anglais de manière impressionniste est peine perdue et relève d'une forme de snobisme et de complexe d'infériorité linguistiques. Il faut penser prioritairement en français. Mais avant de penser en français, il faut se convaincre que le français peut faire face à la musique.

Les actions en faveur de la francisation de l'informatique ont trouvé un nouvel élan lorsque les industries de la langue ont commencé à émerger vers 1984. Les applications industrielles du traitement de la langue, devenue un simple matériau, par les ordinateurs répond à de nombreux besoins dans des secteurs aussi divers que la santé, l'éducation, la gestion, la sécurité, etc., en somme partout où l'informatique peut rendre des services. Si le français n'effectue pas de virage collectif et panfrancophone vers son industrialisation, «ses chances de se maintenir au niveau d'une langue internationale vont en s'amenuisant, du fait de la pression croissante qu'exerceront, dans tout type de communication industrielle, commerciale et scientifique, les ordinateurs capables de manipuler l'anglais, c'est-à-dire la langue des ingénieurs qui construisent actuellement le plus d'ordinateurs» (Hagège 1987:249).

Le contrôle socioprofessionel de toute terminologie ne peut être assuré que par les concepteurs et les usagers de ces vocabulaires. Pour ce qui concerne le français, on a assez répété, et malheureusement pas assez compris ou accepté, que ce ne sont pas les Américains qui vont créer l'informatique en français. Plusieurs groupes professionnels francophones semblent se complaire dans cet état d'attentisme qui se révèle être des plus néfastes puisqu'il ouvre la porte à toutes les excuses et à toutes les exagérations. Le contrôle social de la langue existe puisqu'elle est un bien collectif. «Affirmer que c'est l'usage qui fait la langue apparaît [comme] un faux-fuyant, car la question suivante s'enchaîne : qui fait l'usage?» (Corbeil 1978:158).

Contrôler la langue peut être le début d'un meilleur aménagement économique. On sait que le langage de l'informatique est une institution. Or l'institution est organisée à divers paliers, notamment :

- elle détient un pouvoir d'entraînement considérable;
- elle est un acteur social très important;
- elle peut perturber le projet collectif si elle décide de se marginaliser;
- elle dispose des moyens nécessaires pour définir ses comportements naturels langagiers et pour définir sa norme de référence pour que ses comportements s'y modèlent.

L'orientation interne de la langue est alors le fait de la somme des décisions prises par les individus qui forment le groupe dominant. La conjonction des esprits doit créer la dynamique qui permettra de concevoir que l'informatique en français est un choix éclairé pour une société francophone.

6. La terminologisation

Il ne fait aucun doute que l'informatique est dotée d'un vocabulaire qui dépasse largement plusieurs milliers d'unités. En outre, l'instrumentation lexicographique et terminographique est à peu près au point. Les banques de termes sont saturées, gonflées à bloc par des dizaines de milliers d'unités lexicales de ce domaine. La banque **IBMOT** consigne environ 25 000 termes, le dictionnaire de l'Office de la langue française retrace 12 000 entrées, les dictionnaires classiques se comptent par centaines. Pour le seul sous-domaine de la bureautique, une recherche documentaire a permis de recenser 150 dictionnaires complets ou partiels publiés entre 1980 et maintenant (Cormier et Boulanger

1988). Le noyau dur de la terminologie de l'informatique est presque entièrement répertorié. Il n'exige qu'un entretien périphérique sous la forme de mises à jour régulières afin de réparer quelques errances ou de combler quelques déficits lexicaux constatés dans l'un ou l'autre noeud du super-réseau de la connaissance qu'est l'informatique.

Tout le secteur est terminologisé, consigné, répertorié à l'envi. C'est l'implantation, l'usage qui ne suit pas toujours, parce que la volonté est quasi nulle. La loi du laisser-faire et des voeux pieux l'emporte sur la conviction que le français peut relever le défi de l'informatique. L'idée est si profondément ancrée dans les cerveaux que l'arbre cache la forêt. Il ne faut pas se leurrer non plus; les linguistes ne peuvent à eux seuls imposer l'usage des terminologies et aucune loi incitative n'abattra le mur de la résistance intérieure, pas plus que les termes les plus esthétiques passeront dans le discours des utilisateurs qui en refusent l'emploi.

On laisse l'informatique se terminologiser en anglais en croyant qu'elle se francisera par la suite. C'est là une erreur de jugement fondamentale et la plus belle preuve du transfert et de l'assimilation. A bien examiner la situation de près, on considèrera que la francisation consiste à transiter d'une langue à une autre par le recours aux emprunts, aux calques et, pour une portion minime, à la création interne naturelle. La terminologie française actuelle de l'informatique résulte à peu près intégralement du transfert du savoir-faire, des référents ou des notions élaborés par les Américains. Le matériel linguistique américain est donc adopté, grâce à l'emprunt, et adapté, grâce au calque et à la traduction. Le monopole de l'anglais n'est aucunement brisé si la filiation demeure par trop visible. Il paraît évident que désormais une langue vivante comme le français doit se savoir capable de transformer sa propre actualité en immortalité glorieuse et capitaliser sur son futur sans recourir à la médiation constante d'une langue étrangère et à des modèles technolectaux allogènes ou importés.

7. Conclusion: Rien ne se perd, rien ne se crée

Dans l'introduction, il a été signalé comment tout au long des siècles de l'histoire du français, des spécialistes ont été sensibles aux difficultés, au ressourcement et à l'évolution de leur langue. Seuls les domaines ont varié au gré des besoins des usagers et des découvertes scientifiques ou technologiques nouvelles. Au XVI[e] siècle, les spécialistes avant-gardistes se sont mis à combattre le latin et l'italien, à lutter contre des jargons ecclésiastiques, scolastiques et juridiques. Au XX[e] siècle, on se plaint de l'invasion massive, outrageuse parfois, de l'anglais, on épingle les jargons que sont la plupart des technolectes, et parmi eux l'informatique prend place au premier rang. Le caroussel du temps n'a fait que déplacer les pôles d'attraction et fournir des moyens auxquels les époques passées n'ont pu recourir, et pour cause. Depuis cinq siècles, l'élargissement continu des cadres du lexique général et du lexique terminologique français est un processus réitératif dialectiquement lié aux changements et aux démotivations observés dans les sociétés qui se sont succédées. «C'est là une caractéristique profonde des langues humaines, qui d'usures en réfections, parcourent les voies d'un éternel retour» (Hagège 1983:59).

Il n'est pas plus difficile de créer des mots français en informatique que dans n'importe quelle autre sphère des connaissances. Les médecins et les minéralogistes ont des règles qui fonctionnent et dont ils s'accommodent. Il n'est pas plus difficile de former des néologismes à l'ère de l'exploration spatiale qu'au siècle de Rabelais. En soi, une langue peut tout nommer. «La décision de fonctionner dans sa langue maternelle ou dans une langue «nationale» (en élaborant des terminologies) ou de se résigner à emprunter la langue-outil avec la matière à travailler est politique, car toute langue est capable de tout nommer : l'impression trop fréquente que certaines sémantiques ne peuvent répondre au besoin notionnel relève de l'idéologie» (Rey 1979:66-67).

Dans la francophonie d'aujourd'hui, la francisation et la néologie s'appuient sur des fondements institutionnels solides. La création linguistique est systématisée et organisée rationnellement un peu partout où l'aménagement linguistique est en cours. Les organismes d'intervention sur la langue qui

émanent des politiques linguistiques ont le pouvoir de créer, d'harmoniser et de normaliser des terminologies. Malgré quelques erreurs de parcours observées à l'occasion, ils accomplissent un travail dont les résultats sont fiables et probants. La résistance et les rejets proviennent des milieux socioprofessionnels, des groupes qui sont les premiers producteurs et utilisateurs de termes et qui les répercutent sur l'ensemble du public. Quant à celui-ci, il perpétue dans l'usage ce qu'on lui enseigne et il fait vivre les mots qu'on lui vend avec les produits qu'il acquiert.

Le français devra sa place dans l'espace informatique à l'augmentation des usagers ordinaires et à la vulgarisation de la technique. L'un des fondements du processus de motivation demeure la synergie entre la langue et l'élément humain qui génère la volonté d'agir. Sortir du ghetto terminologique des superspécialistes ne suffit pas. La francomanie de la société moderne doit résolument être associée aux aspirations socio-économiques et volontaristes qui conditionnent le déploiement des communications linguistiques en «langage maternel françois». Tout réexamen du statut du français dans l'informatique est commandé par cette prémisse. Au fond, le problème est relativement simple : ce n'est pas la quantité de français dans la micro-informatique qui est mise en cause, mais bien son statut. Il faut oeuvrer avant tout sur l'aménagement du statut du français en informatique, c'est-à-dire sur sa capacité fonctionnelle de véhiculer la technologie aussi bien dans les hautes sphères spécialisées que dans les conversations courantes. L'ensemble des facteurs de résistance, qu'ils soient d'ordre linguistique, sociolinguistique, psycholinguistique ou extralinguistique, se ramène à l'évaluation du système de la langue et non pas au corpus, c'est-à-dire à la généralisation d'un usage pris comme modèle.

Enfin, il faut bien s'avouer que si l'informatique paraît être un langage, elle ne peut certainement prétendre être une langue. C'est un sphinx qui peut être dompté. Si l'on examine attentivement la petite famille lexicale de départ, on remarque qu'un seul des quatre termes n'a pas été pensé en français. En effet, *ordinateur*, *informatique* et *logiciel* sont des formations autochtones tandis que *artificiel*, malgré son allure française, est calqué sur l'anglais *artificial*. Il est à noter que *ordinateur* et *logiciel* ont évincé des concurrents anglais (*computer* et *software*) en peu de temps; quant à *artificiel*, son intégration est complète, facilitée qu'elle fut par l'existence d'un homonyme français.

Une idée implicite doit s'incarner dans le corps doctrinal de chaque ordre de la connaissance, s'incarner assez profondément pour n'avoir jamais besoin d'être formulée, sinon de biais et comme avec une pudeur révérencielle. L'idée consiste à croire que la vérité du politologue, du médecin, du savant, du technicien, en somme du spécialiste, est incapable de devenir la vérité de tous si elle se contente de demeurer cantonnée dans son exposé technique ou scientifique. Pour atteindre le lieu où elle sera unanimement accueillie, cette vérité doit savoir renoncer à l'orgueil de sa spécialité pour se faire éloquence, c'est-à-dire recevoir l'approbation des usagers conviés au banquet linguistique. Ce qui suppose le choix d'un lexique et de termes attestés par un passé français et trempés dans une matière reconnue par tous. Par delà les formes sélectionnées, le modèle privilégié doit valoriser «une élégance et une clarté heureuse qui rendent la vérité séduisante et transparente à tous, un pathétique et une force d'imagination qui sache faire aller la vérité jusqu'au coeur de tous» (Fumaroli 1986:382). Le reste appartient à l'histoire de demain. Personnellement, je ne doute pas que le français ait une destinée canonique en informatique, capable de lui assurer un avenir prometteur à l'enseigne de ce que lui souhaitait déjà Ronsard à la Renaissance, qui fut l'une des grandes époques fastes de l'incubation de la langue française. Laissons au français contemporain le temps d'accomplir son oeuvre dans chaque sphère du savoir.

BIBLIOGRAPHIE

BOULANGER (Jean-Claude), 1988a, Un miroir aux alouettes: le calque en intelligence artificielle, dans *Le Langage et l'homme*, vol. 23, fasc. 1, n° 66, p. 3-13.

____ 1988b, *L'évolution du concept de «néologie» de la linguistique aux industries de la langue*, communication présentée au Colloque sur l'histoire de la terminologie, Bruxelles, 25 et 26 mars 1988, 23 p. dactylographiées, à paraître dans les actes.

CORBEIL (Jean-Claude), 1978, Les conditions de succès des lois à caractère linguistique, dans *Les implications linguistiques de l'intervention juridique de l'État dans le domaine de la langue. Actes du Colloque international de sociolinguistique*, Lac-Delage (Québec), 3 au 6 octobre 1976, Québec, Office de la langue française, Éditeur officiel du Québec, mars, p. 155-164.

CORMIER (Monique) et Boulanger (Jean-Claude), 1988, (sous la direction de), *Analyse des répertoires terminologiques de la bureautique et de quelques domaines connexes. Bibliographie sélective et critique*, Édition provisoire, Laval (Québec), Centre canadien de recherche en informatisation du travail, 152 p.

FUMAROLI (Marc), 1986, La coupole, dans *Les lieux de mémoire. II. La nation*, coll. «Bibliothèque illustrée des histoires», Paris, Éditions Gallimard, vol. 3, p. 321-388.

GOUJON (Pierre), 1985, L'informatique individuelle, dans *Universalia 1985. Les événements, les hommes, les problèmes en 1984*, Paris, Encyclopaedia Universalis France, p. 400-403.

HAGÈGE (Claude), 1983, Voies et destins de l'action humaine sur les langues, dans *La réforme des langues. Histoire et avenir*, vol. 1, Hambourg, Buske Verlag Hamburg, p. 11-68.

____ 1987, *Le français et les siècles*, Paris, Éditions Odile Jacob, 272 p.

HUMBLEY (John), 1987, L'emprunt sémantique dans la terminologie de l'informatique, dans *La fertilisation terminologique dans les langues romanes. Actes du colloque de Paris*, octobre 1986, publiés dans *Meta*, vol. 32, n° 3, septembre, p. 321-325.

MARCELLESI (Christiane), 1973, Le langage des techniciens de l'informatique: quelques aspects de leur vocabulaire écrit et oral, dans *Les vocabulaires techniques et scientifiques*, publiés dans *Langue française*, n° 17, février 1973, p. 59-71.

____ 1979, Retour aux sources : quelques aspects du vocabulaire de l'informatique, dans *Néologie et lexicologie. Hommage à Louis Guilbert*, coll. «Langue et langage», Paris, Librairie Larousse, p. 176-183.

MENEY (Lionel), Juhel (Denis) et Mepham (Michael), 1987, *Le guide de l'utilisateur (IBM PC): Difficultés linguistiques et rédactionnelles*, coll. «K», n° 2, Québec, Université Laval, Centre international de recherche sur le bilinguisme, 18 p.

QUEMADA (Bernard), 1978, Technique et langage, dans *Histoire des techniques. Techniques et civilisations. Techniques et sciences*, coll. «Encyclopédie de la Pléiade», n° 41, Paris, Éditions Gallimard, p. 1146-1240.

Rapport d'étape du comité Canada-Québec sur le développement du logiciel d'expression française et faits saillants, s.l., Ministère des Communications du Canada et Ministère des communications du Québec, juin, 6 p.

REY (Alain), 1979, *La terminologie: noms et notions*, coll. «Que sais-je», n° 1780, Paris, Presses universitaires de France, 128 p.

TURCOTTE (Roselyne), 1988, *Vers une terminologie des arts infographiques*, Rapport sur le projet de recherche intitulé «Terminologie, néologie et informatisation dans un contexte bilingue canadien», sous la direction de Jean-Claude Boulanger et Monique Cormier, Université Laval, CEFRIO, 120 p., à paraître.

MESURE DU DEGRÉ DE DIFFICULTÉ DE TEXTES RELATIF À L'INFORMATIQUE[1]

Martine Bourque
Langues et linguistique
Université Laval

Cette étude a pour objectif de créer un outil d'aide à l'évaluation de textes dans le domaine de l'informatique, du point de vue du degré de difficulté du vocabulaire. Le résultat d'une telle étude serait destiné à un éventuel lecteur voulant sélectionner des textes adaptés à son niveau de compréhension et à son expérience. L'utilisation d'un indice de difficulté devrait permettre de distinguer les textes de vulgarisation des textes techniques.

Un tel outil doit être partiellement automatisé pour faciliter la tâche des évaluateurs et augmenter ainsi le nombre de textes pouvant être évalués.

Jusqu'à maintenant, peu de recherches ont été effectuées sur le degré de difficulté des textes, du moins en ce qui concerne la difficulté du vocabulaire. L'évaluation de la difficulté des textes était plutôt basée sur la complexité des structures de phrases et la densité de l'information contenue.

Cependant, trois domaines de recherche, sans toucher directement le sujet de cette étude, abordent la notion de difficulté du mot et peuvent guider nos recherches.

D'abord, l'enseignement du vocabulaire en langue seconde, surtout depuis les années 1960, a donné lieu à une classification du vocabulaire en fonction de la difficulté d'apprentissage. Quelques-uns des critères utilisés pour cette classification peuvent s'intégrer dans la création d'un indice de difficulté lexicale tandis que d'autres n'apportent aucune indication utile. Nous pensons, dans ce dernier groupe, à la fréquence d'apparition des mots dans le discours.

Dans le domaine de la statistique lexicale, vers la fin des années 1970, des critères de classification des mots ont été émis. Certains d'entre eux, en particulier la catégorie grammaticale, le mode de formation et la complexité graphique, sont susceptibles d'influencer le degré de difficulté du vocabulaire.

Enfin, la psychologie du langage s'est penchée, vers 1980, sur la compréhension du mot. Ces recherches ont fait ressortir des éléments influençant le degré de difficulté. Il s'agit de la longueur, du degré d'abstraction et du degré de spécificité. On dit qu'un terme a un haut degré de spécificité lorsqu'il n'est pas général, qu'il a un contexte d'utilisation restreint, c'est-à-dire qu'il est adapté à une situation donnée, et n'est pas facilement transportable.

Pour résumer, les recherches disponibles actuellement ne portent pas spécifiquement sur le degré de difficulté des textes en fonction du vocabulaire. Nous en sommes encore au stade de la recherche d'une méthode d'évaluation.

---

[1]Communication présentée lors du 56e Congrès de l'ACFAS, Moncton, 9-13 mai 1988.

Il paraissait intéressant, pour tester la méthode préconisée, de comparer deux textes du point de vue du degré de difficulté. Idéalement, ces deux textes devaient traiter du même sujet mais à des degrés de difficulté présumés différents. Notre choix s'est porté sur le *Guide de l'usager* et le *Manuel de référence* du système d'exploitation Microsoft DOS.

L'hypothèse avancée veut que le *Manuel de référence* présente de plus grandes difficultés que le *Guide de l'usager*, comme c'est habituellement le cas pour n'importe quelle documentation se rapportant à un logiciel.

La première étape de l'analyse consiste à saisir le texte au moyen d'un éditeur de texte conventionnel et comporte une recherche des syntagmes. En effet, on sait que le vocabulaire technique, et plus spécialement ici le vocabulaire de l'informatique, recèle de nombreux syntagmes. La définition de *syntagme* est celle du «syntagme terminologique», à savoir une suite de mots qui réfère à une notion, sans que ces mots soient unis par un trait d'union (ex. : «chemin de fer», dans la langue générale et «unité de disquette», dans le vocabulaire de l'informatique).

Les éléments de syntagmes doivent être regroupés avant de procéder à l'analyse du vocabulaire. Les syntagmes du vocabulaire spécialisé sont repérés manuellement et les espaces entre les mots qui les constituent sont remplacés par des caractères quelconques, en l'occurrence des traits de soulignement.

La segmentation en mots peut ensuite être effectuée grâce à un programme qui sépare le texte aux espaces et aux ponctuations. Les syntagmes, qui ne sont plus entrecoupés d'espaces, conservent leur unité.

Le fichier obtenu après la segmentation contient tous les mots du texte initial, y compris le vocabulaire de la langue générale. Les mots qui appartiennent au vocabulaire de l'informatique doivent être repérés et marqués pour constituer une liste des mots à analyser.

La deuxième étape comprend le choix et la pondération des variables. Elle débute par la constitution d'une liste, aussi exhaustive que possible, des variables susceptibles d'influencer le degré de difficulté. Il n'y a aucune sélection à ce niveau : toutes les variables potentielles sont conservées.

Ces variables sont classées en deux catégories, soit les variables lexicales et les variables contextuelles. Le premier groupe comprend la catégorie grammaticale, la longueur du terme, le degré d'abstraction, l'origine ou mode de formation, le degré de spécificité, la complexité graphique et la motivité, cette dernière se définissant par la propriété d'un terme de représenter les signes distinctifs essentiels de l'objet dont il désigne le concept. Les variables contextuelles servent à indiquer la présence d'éléments susceptibles d'expliquer un terme. Ce sont la définition, l'exemple, l'illustration, le contexte encyclopédique, les synonymes et les collocations.

La méthode pour la sélection et la pondération des variables sélectionnées a pour point de départ une grille dont un exemple simplifié apparaît à la figure 1.

FIGURE 1:

	variable 1	variable 2	variable 3	degré de difficulté
terme 1	1	4	6	2
terme 2	8	2	0	9
terme 3	3	3	2	7
terme 4	1	1	7	5

La grille de la figure 1 étant un modèle simplifié, elle ne comporte ni toutes les variables à analyser, ni un nombre de termes suffisant pour obtenir des résultats probants. De même, les valeurs numériques sont aléatoires et fournies simplement à titre indicatif.

La grille est composée d'une colonne par variable et d'une ligne par terme, chacune de ces lignes contenant les valeurs associées à chaque variable pour ce terme. Il n'est pas nécessaire, pour cette étape qui vise à choisir et pondérer les variables, d'inclure tous les termes spécialisés du texte.

La dernière colonne sert à indiquer un degré de difficulté intuitif pour chaque terme. Intuitif signifie que ce degré n'est basé sur aucun indice quantitatif, mais est plutôt issu d'une impression personnelle de la difficulté observée à la lecture.

A partir de cette grille et au moyen d'un logiciel de statistiques, une analyse multivariée permet de déterminer quelles variables ont influencé le degré de difficulté et de mesurer l'indépendance réciproque des différentes variables de façon à pouvoir éliminer une variable qui aurait exactement la même influence qu'une autre sur l'indice de difficulté.

L'analyse multivariée révèle aussi dans quelles proportions chaque variable influence le degré de difficulté. Par exemple, il se peut que la variable 1 contribue à trente pour cent dans l'indice de difficulté alors que la variable 2 n'influence que dans une proportion de dix pour cent.

Pour que cette technique de sélection soit validée, il faudrait que la cote de difficulté choisie ici intuitivement, s'appuie plutôt sur une enquête auprès des usagers. Une telle validation implique l'élaboration d'un questionnaire demandant à des lecteurs d'indiquer, pour chaque terme, une cote de difficulté sur une échelle de 1 à 10, et de faire parvenir ce questionnaire à un nombre relativement élevé de personnes jugées représentatives de la population cible. Le travail découlant de cette enquête dépasse le cadre de la présente recherche. Nous nous en tiendrons donc à un schéma théorique qui devra ultérieurement être validé.

La troisième étape consiste, pour chaque terme, à part ceux qui ont été utilisés dans la grille d'analyse et pour lesquels le travail a déjà été fait, à associer une valeur pour toutes les variables retenues. Il n'est plus nécessaire d'associer un degré de difficulté intuitif puisque ce procédé n'était utilisé que pour l'évaluation des variables. Les indices des autres termes sont calculés automatiquement.

Le calcul des indices de difficulté fait l'objet de la quatrième étape. Il s'effectue selon la méthode suivante : pour chaque terme, l'indice de difficulté est obtenu en multipliant la valeur de chaque variable par sa pondération (liée au degré d'influence), puis en additionnant toutes les valeurs pondérées. Le total est ramené sur 10.

Une moyenne des cotes de difficulté est ensuite calculée pour l'ensemble du texte. Les degrés de difficulté minimum et maximum rencontrés sont également recueillis.

Au cours de la cinquième étape, les résultats obtenus à partir des deux textes du corpus sont comparés pour vérifier l'hypothèse de départ, à savoir que le manuel de référence doit présenter un degré de difficulté plus élevé que le guide de l'usager.

Une méthode d'analyse de la difficulté des textes basée essentiellement sur le vocabulaire offre l'avantage d'aborder un nouvel aspect jusqu'à présent très peu exploré.

En ce qui a trait aux variables liées au terme indépendamment du contexte, autrement dit les variables lexicales, une telle analyse permet de n'effectuer qu'une seule saisie des valeurs de variables pour chaque terme spécialisé. Les mêmes valeurs sont automatiquement associées aux autres occurrences de ces termes et peuvent être conservées dans un index pour s'appliquer à d'autres textes.

Les programmes requis pour cette étude exécutent principalement six tâches : l'édition de texte, la segmentation, la sélection du vocabulaire à traiter, la saisie des variables lexicales, la saisie des variables contextuelles et les fonctions statistiques (choix et pondération des variables, calcul de l'indice de difficulté).

Parmi ces programmes, certains doivent être en mesure de traiter des difficultés particulières.

Nous avons déjà fait mention des contraintes liées à l'édition du texte, à sa segmentation et à la sélection du vocabulaire. Quant aux fonctions du logiciel de statistiques, elles ne posent pas de problème particulier.

Il reste donc à considérer les programmes de saisie des variables lexicales et contextuelles.

En ce qui concerne les variables lexicales, il n'est pas nécessaire de conserver une liste de toutes les occurrences présentes dans le texte puisque les valeurs des variables sont valables pour les différentes occurrences d'un même terme. Le programme doit donc construire une liste qui élimine les mots repris dans le texte. Cette fonction, bien que simple en apparence, exige que le texte soit préalablement lemmatisé (les mots sont tous réduits à leur forme de base) sinon l'ordinateur considèrera comme différents les mots *disquette* et *disquettes* qui possèdent cependant les mêmes valeurs pour les variables lexicales.

Pour les variables contextuelles, les valeurs varient d'une occurrence à l'autre. Par exemple, il se peut que la première occurrence d'un terme ne soit pas accompagnée d'un exemple alors que ce dernier apparaît avec la deuxième occurrence.

D'ailleurs, à partir du moment où l'occurrence d'un terme est définie, les occurrences suivantes doivent être aussi considérées comme étant définies, même si la définition n'est pas répétée dans le texte.

Il convient donc que cette valeur soit enregistrée automatiquement pour toutes les occurrences qui suivent un mot défini, évitant ainsi la saisie manuelle d'une même valeur à plusieurs reprises.

Le programme doit aussi être capable de retrouver un contexte pour le présenter avec une occurrence afin de vérifier la présence d'un exemple, d'une définition, d'une illustration, etc., ceci pour permettre la saisie des valeurs des variables contextuelles.

Pour consigner les valeurs des deux types de variables, il faut faire appel à deux fichiers : un fichier de lemmes avec les valeurs des variables lexicales et un fichier des occurrences accompagnées des valeurs des variables contextuelles. Le fichier des lemmes pourra servir ultérieurement pour d'autres textes. Son volume augmente au fur et à mesure que des termes nouveaux apparaissent dans un texte. Par contre le fichier des occurrences ne vaut que pour un seul texte.

En conclusion, il serait intéressant de tester la méthode préconisée ici sur un texte complet d'abord, puis sur des portions de texte de différentes longueurs et recueillies à différents endroits dans le texte entier pour déterminer la portion de texte la plus petite dont l'analyse donnera le même

résultat que celle du texte entier. Cela permettrait d'évaluer des textes sans en saisir le contenu entier, sans quoi la méthode ne serait pas appliquable à un nombre suffisant de textes pour être utile à des usagers.

Des cotes basées sur les variables lexicales seulement pourraient aussi être mises à la disposition d'auteurs de manuels d'informatique afin de leur suggérer un vocabulaire adapté à la clientèle visée. Par exemple, l'auteur d'un livre d'initiation à l'informatique pour enfants pourra éviter les termes présentant une cote élevée ou, s'il doit les utiliser, veillera à les définir adéquatement.

Pour conclure, nous rappellerons que l'analyse proposée ici en est encore au stade expérimental et que des recherches plus poussées seraient nécessaires pour la valider. Il reste donc beaucoup de travail à faire pour obtenir une mesure d'évaluation du degré de difficulté des textes en se basant sur le vocabulaire.

BIBLIOGRAPHIE

HAMMERLY, Hector. *Synthesis in Second Language Teaching. An Introduction to Linguistics.* Preliminary Edition. Second Language Publications. Blaine. Wash. 1982. 693 p.

HENDERSON (Leslie). *Orthography and Word Recognition in Reading.* Academic Press. London. 1982. 397 p.

MACKEY (William Francis). *Language Teaching Analysis.* Longman Group Ltd. London. 1965. 562 p.

MULLER (Charles). *Principes et méthodes de la statistique lexicale.* Collection «Langue Linguistique Communication». Classiques Hachette. Paris. 1977. 206 p.

PHAL (André). «Les groupes de mots et les problèmes qu'ils posent dans la préédition de textes scientifiques destinés à l'analyse mécanographique», dans *Cahiers de lexicologie*, vol. IV, n° 1, 1964. pp. 45-60.

L'INFORMATIQUE EN FRANÇAIS: AUJOURD'HUI ET DEMAIN[1]

Pierre-Étienne Laporte
Office de la langue française

Il sera question dans les pages qui suivent de l'informatique en français. Sujet qui est également de la plus grande importance pour l'organisme que je préside, compte tenu de ses enjeux énormes pour la francisation et, en particulier, pour la francisation des entreprises.

La Charte de la langue française, sanctionnée en 1977, concerne non seulement le statut du français comme langue officielle, en particulier son statut juridique, mais aussi son statut d'usage dans les principaux domaines de la vie publique : le travail, la consommation des biens et services ainsi que l'enseignement.

C'est à des organismes gouvernementaux qu'a été confiée l'application de la Charte, qui vise à faire du français la langue d'utilisation généralisée au Québec.

L'Office de la langue française, qui fête son vingt-cinquième anniversaire cette année, a reçu le mandat de définir et conduire la politique québécoise en matière de recherche linguistique et terminologique et de veiller à ce que le français devienne, le plus tôt possible, la langue des communications, du travail, du commerce et des affaires dans les organismes de l'Administration et dans les entreprises.

Le Conseil de la langue française doit conseiller la ministre responsable de l'application de la Charte sur la politique québécoise de la langue française et sur toute question relative à l'interprétation et à l'application de la Charte.

La Commission de protection de la langue française a été instituée pour traiter des questions se rapportant au défaut de respect de la Charte. Quant à la Commission de toponymie, elle a compétence pour établir les critères de choix et les règles d'écriture de tous les noms de lieux.

Dans le cadre de son mandat, l'Office de la langue française a dû s'intéresser à la nouvelle technologie qu'est l'informatique lorsqu'on a constaté que d'importants éléments de gestion de l'entreprise l'utilisaient de plus en plus.

En effet, jusqu'en 1979, l'informatique était surtout utilisée par les grandes entreprises qui avaient les ressources nécessaires pour acquérir de grosses machines, de l'équipement et du personnel spécialisés.

---

[1]Allocution prononcée au Palais des Congrès de Montréal, le 6 octobre 1987.

Cependant, dans les années 1980, nous avons été témoins de l'avènement de la micro-informatique et de l'arrivée de nouveaux logiciels comme les traitements de texte et les tableurs électroniques. Mais surtout, nous avons assisté à l'entrée de l'informatique dans la petite et moyenne entreprise; puisqu'il devenait possible d'avoir accès à la gestion informatisée à un coût abordable.

Ces logiciels provenaient des États-Unis, étaient vendus en anglais aux entreprises et, souvent, ne pouvaient être modifiés.

Utilisant des documents remis aux consommateurs au moment d'une transaction, tels que les factures, les reçus ou les modes d'emploi, bien des entreprises ont anglicisé ce domaine à la suite de l'arrivée de l'informatique. De plus, les documents administratifs internes sont, malheureusement, souvent passés du français à l'anglais à cause de l'informatique. Nous assistions à un retour à l'anglais par l'introduction de l'informatique.

Toutefois, il était possible d'obtenir certains logiciels en français, ou de modifier, traduire et adapter certains logiciels de langue anglaise. La demande pour ces produits s'accroissant, nous avons vu apparaître des logiciels de traitement de texte capables de reproduire les signes diacritiques du français, des logiciels de gestion complètement francisés, ce qui témoignait d'une prise de conscience de la spécificité du marché québécois et d'une adaptation à la clientèle francophone.

Confronté à l'introduction de l'informatique de façon massive au sein des entreprises et des organismes gouvernementaux, l'Office a, en 1983, chargé un comité d'étudier la question de la francisation de l'informatique. Ce comité a proposé une série de recommandations qui ont été approuvées par le Conseil d'administration de l'Office en octobre 1987.

Ces recommandations furent alors envoyées aux entreprises utilisant l'informatique, aux fabricants et aux fournisseurs de produits informatiques ainsi qu'aux organismes gouvernementaux.

Selon les recommandations entérinées par l'Office, devraient être disponibles en français:

Pour ce qui est de la *documentation*:
- la documentation publicitaire;
- les documents relatifs au matériel et aux logiciels;
- les cours de formation et le matériel didactique.

Pour ce qui est du *matériel*:
- les inscriptions sur le matériel, sur les boutons de commande et sur les touches de clavier;
- les accents français.

Pour ce qui est des *logiciels*:
- les menus, les commandes et les messages affichés à l'écran ainsi que les textes affichés à l'écran ou imprimés.

Ces propositions, faites aux 370 000 entreprises établies au Québec et aux organismes de l'Administration, tenaient compte à notre avis des possibilités actuelles de l'informatique ainsi que de l'évolution prévisible de la technologie.

Il faut souligner que l'Office n'exige pas la francisation des langages de programmation ni la francisation des documents hautement techniques, souvent périmés dès leur impression. L'Office désire, cependant, que les utilisateurs puissent travailler avec des outils en français et que les consommateurs puissent être servis en français. Ce qui, faut-il le mentionner, correspond aux attentes des uns et des autres dans la grande majorité des cas.

Il convient d'ajouter que ces propositions sont incluses dans les programmes de francisation qui touchent les entreprises employant 50 personnes et plus, ainsi que dans les programmes qui touchent les organismes de l'Administration. Ces entreprises et ces organismes, dont l'Office gère la francisation, représentent 54% de la population active du Québec.

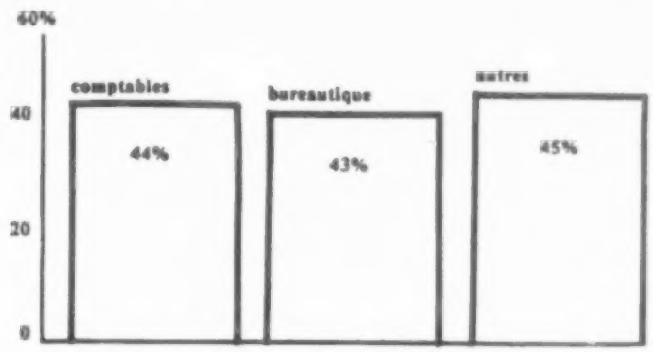
L'Office constate aujourd'hui, avec satisfaction, que ses propositions étaient réalistes. En effet, nous trouvons un grand nombre de logiciels en français possédant, d'ailleurs, tous les signes diacritiques de la langue française.

Une étude de 1986, effectuée par le Centre de recherche en gestion de l'Université du Québec à Montréal, intitulée «Le dirigeant québécois face à l'informatique» et à laquelle 1 725 dirigeants ont choisi de participer confirme que, dans une proportion importante, les entreprises québécoises se sont informatisées en français.

Nous mentionnons ci-après quelques résultats de cette étude. Le premier graphique donne les pourcentages d'utilisation des logiciels français par les entreprises dans trois secteurs d'activité.

FIGURE 1:

Pourcentage des entreprises utilisant des logiciels français



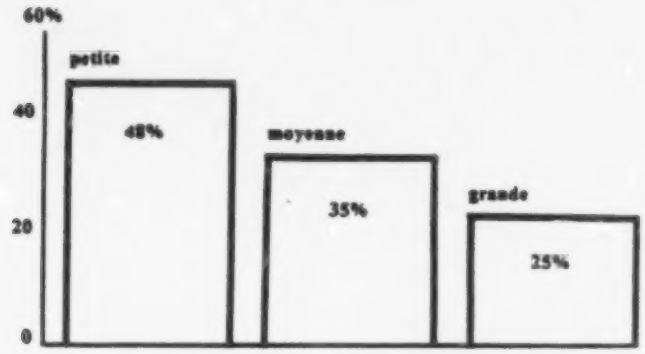
Source: *Le dirigeant québécois face à l'informatique - 1987.*

On constate qu'environ 45% des entreprises utilisent des logiciels comptables, de bureautique et de gestion en français.

Les trois autres graphiques fournissent les pourcentages d'utilisation des logiciels en français selon la taille de l'entreprise : logiciels comptables, de bureautique et de gestion.

FIGURE 2:

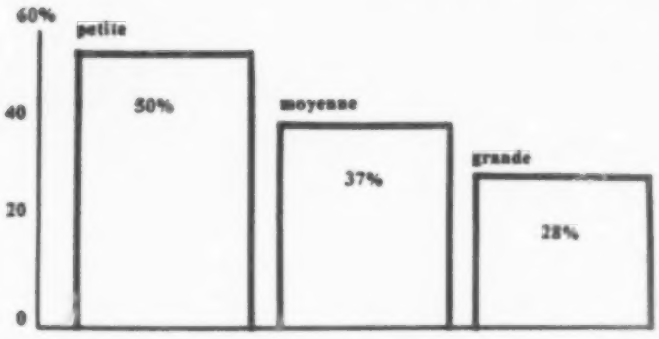
Logiciels de bureautique utilisés en français selon la taille de l'entreprise



Source: *Le dirigeant québécois face à l'informatique - 1987.*

FIGURE 3:

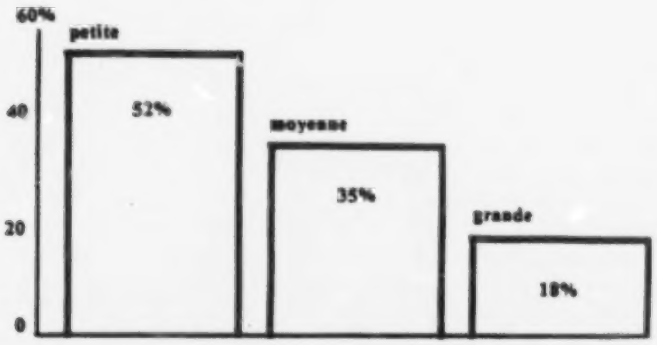
Logiciels comptables en français selon la taille de l'entreprise



Source: *Le dirigeant québécois face à l'informatique - 1987.*

FIGURE 4:

Logiciels de gestion en français selon la taille de l'entreprise



Source: *Le dirigeant québécois face à l'informatique - 1987.*

La taille des entreprises, ainsi qu'on a pu le constater, influence grandement le choix de la langue des logiciels : les petites entreprises sont de beaucoup les plus francisées, suivies des entreprises de taille moyenne. Par contre, le degré de francisation des grandes entreprises, pour ce qui est du secteur informatique, est significativement inférieur à celui des petites et moyennes entreprises.

Ces différences s'expliquent, selon les chercheurs, par le fait que les grandes entreprises se sont informatisées à l'époque où n'étaient disponibles que les logiciels en anglais. Les petites et moyennes entreprises s'étant informatisées plus tardivement, elles ont eu accès à des produits informatiques en français.

Il faut également mentionner que les grandes entreprises canadiennes et nord-américaines ont beaucoup plus tendance à utiliser l'anglais. Toutefois, nous décelons au sein des grandes entreprises une forte tendance à l'informatique en français, tendance qui devrait se consolider dans l'avenir, compte tenu des pressions de l'opinion publique et du marché parallèlement à l'évolution prévisible de la technologie.

Il faut ajouter que l'étude mentionnée plus haut démontre que beaucoup d'entreprises veulent travailler en français et utilisent des logiciels en français lorsque ces derniers sont disponibles sur le marché. De plus, ces entreprises constituent une clientèle énorme pour des logiciels en français au Québec où, il faut le rappeler, il y a 366 000 entreprises qui emploient moins de cinquante salariés. En outre, beaucoup de grandes entreprises songent actuellement à modifier leur gestion informatique pour tenir compte de leur politique linguistique : ce qui était uniquement en anglais sera francisé ou bilingue.

Ce qui vient d'être souligné ne doit pas donner l'impression que nous ignorons les difficultés de l'implantation de l'informatique en français.

Cependant, sans entrer dans les problèmes techniques qui sont du ressort des informaticiens, nous pouvons affirmer que des solutions intéressantes ont été trouvées et que des chercheurs, tant du secteur privé que du secteur public, étudient la question.

Un grand nombre d'entreprises offrent déjà aux utilisateurs québécois du matériel et des logiciels de traitement de texte et de gestion en français, accompagnés de manuels d'utilisation en français.

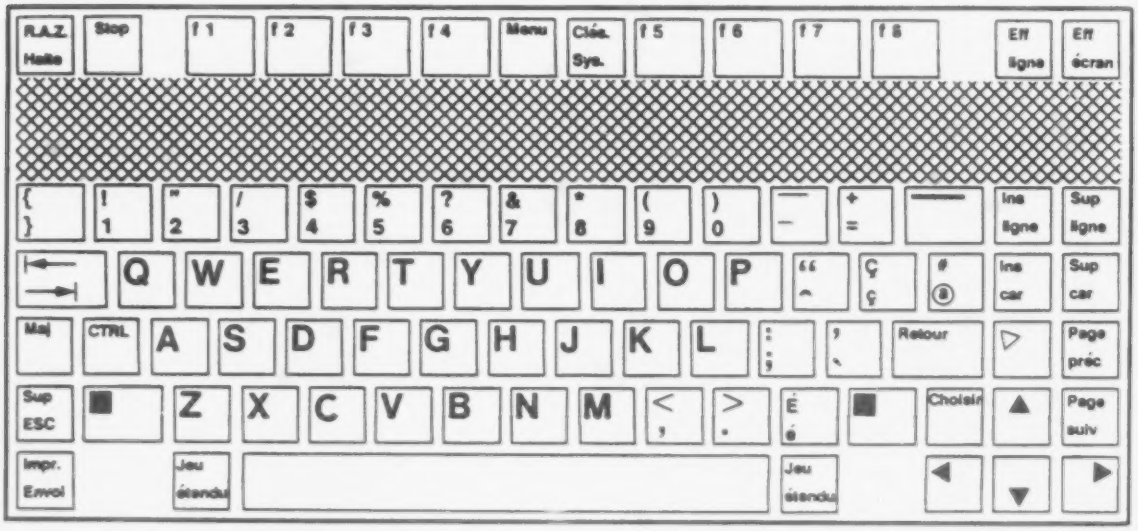
La majorité des logiciels conçus au Québec et adaptés aux entreprises québécoises sont disponibles en français.

On ne remet plus en question aujourd'hui la capacité des Québécois de créer des logiciels de langue française; leur rentabilité fait même partie de la politique de mise en marché de grandes entreprises.

L'exigence de convivialité et, de plus en plus, les attentes des clients demandent une adaptation à la langue de l'utilisateur.

Le matériel informatique est maintenant francisé : les claviers sont francisés chez IBM, Wang, Digital, HewlettPackard, Apple, AES Data, Northern Telecom. Ils le sont également dans d'autres entreprises qui utilisent, pour certains de leurs appareils, ce qui est communément appelé, le clavier canadien-français, dont voici une illustration :

CLAVIER CANADIEN-FRANÇAIS



Il faut souligner que les touches de fonction du clavier ont été francisées et qu'il est possible d'obtenir les accents français directement.

Des études et des travaux sont actuellement en cours pour trouver des solutions à long terme aux problèmes que pose la francisation des logiciels. Parmi les études les plus intéressantes publiées récemment, il faut signaler celles de Monsieur Alain Labonté, du Bureau central de l'informatique du ministère des Communications du Québec. La première étude traite des aspects techniques de l'utilisation du français en bureautique. La seconde étude, réalisée dans le cadre des travaux du comité Canada-Québec sur le développement du logiciel d'expression française, traite de l'apport du multilinguisme lors de la mise au point technique des logiciels.

Ces deux études, dont la lecture est fortement recommandée, présentent clairement des solutions à l'adaptation des logiciels aux différentes langues nationales et suggèrent la conception de logiciels à portée multilingue. En plus d'être à l'avant-garde de la réflexion dans ce domaine, ces études laissent entrevoir des perspectives fort encourageantes quant à la coopération possible à l'échelle de la francophonie internationale.

Il faut signaler également que la Centrale des bibliothèques a répertorié les logiciels conçus au Québec à l'intérieur d'une base de données appelée **LOGIBASE**. Celle-ci vise à répertorier tous les logiciels québécois disponibles sur le marché pour tous les types et modèles d'ordinateurs.

En plus de collaborer avec les organismes gouvernementaux qui s'intéressent à l'informatique en français, l'Office fait la promotion des logiciels et du matériel informatique en français auprès des entreprises québécoises employant 50 employés ou plus.

La Direction des programmes de francisation a reçu le mandat d'aider les entreprises ainsi que les organismes de l'Administration à généraliser l'utilisation du français. Quatre mille entreprises et organismes sont donc invités à utiliser des logiciels en français, ce qui peut être un incitatif puissant pour créer une demande de logiciels et de matériel francisés.

L'Office a également publié une série de lexiques et de vocabulaires susceptibles d'aider les traducteurs, concepteurs de logiciels ou fabricants de matériel à utiliser une terminologie française. Mentionnons, entre autres, le *Vocabulaire du micro-ordinateur*, le *Vocabulaire du logiciel* et le *Vocabulaire du traitement de texte*.

En plus de publier des vocabulaires, l'Office de la langue française permet l'accès à sa banque de données, la **Banque de terminologie du Québec**. Créée en 1974, la banque de l'Office contient aujourd'hui près de trois millions de termes et peut être consultée de deux façons. Les entreprises qui ont des besoins terminologiques importants et qui possèdent le matériel informatique nécessaire peuvent y avoir accès directement par l'intermédiaire des réseaux **Datapac** et **Inet 2000**.

Les autres clients peuvent consulter la Banque de terminologie de l'Office par l'intermédiaire d'un service d'accès en différé. Nos terminologues font alors des recherches dont les résultats sont transmis aux usagers.

Ces études et ces travaux dénotent bien notre volonté d'offrir à nos clientèles des produits adaptés à leurs besoins.

Insistons sur le fait que l'Office, dans ce domaine de son activité, travaille toujours avec l'étroite collaboration des organismes et entreprises spécialisés, et non pas en vase clos.

En février 1986, les ministères des Communications du Canada et du Québec formaient un comité sur le développement du logiciel d'expression française. Le comité avait pour mandat d'examiner l'ensemble des problèmes que vit l'industrie du logiciel d'expression française et de formuler des recommandations quant à l'avenir de cette industrie.

Au début de 1987, la Fédération d'informatique du Québec et l'Office ont participé à la validation de ces recommandations. Les plus significatives d'entre elles sont les suivantes :

Multilinguisme

1- Encourager la création et la commercialisation de logiciels en français à portée multilingue, conçus de manière à pouvoir incorporer les caractères fondamentaux d'une ou plusieurs langues.

Aspects techniques du français

2- Informer l'entreprise et lui offrir toute autre forme de soutien quant aux opérations techniques exigées par le traitement du français en informatique.

Conformité aux normes et usages

3- Informer l'entreprise des travaux de normalisation en cours sur la scène nationale ou internationale, ayant une portée linguistique ou culturelle.

Traduction

4- Encourager la recherche appliquée dans les domaines suivants : la traduction et l'adaptation de logiciels à portée multilingue (ex. traitement de texte), de logiciels de traduction assistée et d'outils d'aide à la traduction.

Coopération avec la francophonie

5- Favoriser les projets de logiciels en français qui nécessitent la coopération des pays francophones.

Rôle des gouvernements

6- Faire usage du pouvoir d'achat des gouvernements et de son effet d'entraînement sur l'industrie pour exiger des logiciels en français à portée multilingue, de conception canadienne et conformes aux normes reconnues quant au traitement du français.

Ces recommandations ont pour objectif le développement des produits conçus par nos entreprises, (cette activité étant créatrice d'emploi) et facilement exportables sur le marché mondial.

Le comité souligne qu'il est urgent que l'industrie québécoise tire profit du caractère linguistique particulier du Québec en Amérique du Nord et développe, pour l'exportation, des logiciels utilisables aussi bien en français qu'en d'autres langues.

Il semble urgent de concrétiser ces recommandations et de profiter du fait que les deux gouvernements consentent à travailler à un projet commun, soit le développement de l'industrie informatique.

Il est aussi impératif, après avoir fait le tour de la question des possibilités de francisation et s'être rendu compte que tout n'a pas été fait, d'admettre qu'il reste encore d'autres défis à relever.

Ainsi, en ce qui a trait à la langue, il va falloir continuer les travaux de terminologie et de normalisation avec la francophonie, et faire la promotion et la diffusion des termes auprès des utilisateurs. L'Office peut s'engager à poursuivre la collaboration avec les entreprises d'informatique et à rester à l'écoute des préoccupations de cette industrie.

Il faudra notamment trouver des solutions concernant la compatibilité des différents langages et matériels et s'entendre sur une norme relative à la position des signes orthographiques sur les claviers d'ordinateurs.

Les utilisateurs doivent pouvoir utiliser leur propre langue et, pour ce faire, la conception de logiciels à portée multilingue devra être généralisée. Les grandes entreprises américaines comme **Microsoft, IBM, Digital** tiennent maintenant compte des différentes langues nationales. La conception de ces logiciels à portée multilingue pourra permettre à nos entreprises d'informatique de servir le marché québécois tout en espérant atteindre le marché mondial.

Comme pour n'importe quel produit ou service, les concepteurs devront se tenir à l'écoute de leur clientèle pour faire en sorte que les nouveaux logiciels soient plus performants, plus faciles d'utilisation et que leur prix de vente soit le plus compétitif possible.

L'Office a l'intention de maintenir ses demandes en ce qui a trait à la francisation de l'informatique tout en tenant compte des difficultés inhérentes à cet objectif.

Dans cet esprit, nous voulons continuer à travailler en collaboration avec les différents paliers de gouvernement afin de favoriser l'achat et l'utilisation de logiciels en français. Nous voulons également collaborer avec les entreprises d'informatique afin de continuer à produire des lexiques et des vocabulaires susceptibles d'être acceptés et utilisés. Nous voulons aussi continuer à examiner, avec les différents intervenants, les solutions techniques acceptables pour la francisation des entreprises. Enfin, nous voulons intervenir auprès des organismes de normalisation.

En contrepartie, nous souhaitons que les entreprises d'informatique donnent satisfaction aux marchés francophones québécois et canadien. Ainsi, les *concepteurs et les fabricants* devraient adapter leurs produits aux milieux auxquels ils sont destinés, et ils devraient rendre simultanément disponibles sur le marché les versions anglaises et françaises de ces logiciels. Les *consultants et les conseillers* devraient informer leurs clients des caractéristiques linguistiques de leurs produits. Les *distributeurs* devraient offrir au comptoir l'option française des produits.

Nous sommes persuadés qu'il existe une réelle demande pour des produits informatiques en français et que cette demande ne fera qu'augmenter.

Pour conclure, rappelons que l'Office est au service de la communauté et veut collaborer à l'atteinte de l'objectif que nous partageons : le développement au Québec du secteur informatique .... en français.

LA LANGUE FRANÇAISE: PROBLÈMES D'INITIATION À L'INFORMATIQUE[1]

Michaël Mepham
Langues et linguistique
Université Laval

Cette communication porte sur les problèmes de francisation tels que perçus avec les oeillères de l'expérience et de la réflexion personnelles, acquises sur une période d'un quart de siècle en qualité d'utilisateur et d'enseignant de l'informatique linguistique à l'université Laval.

1. Problématique et enjeux

Vaut-il vraiment la peine de franciser l'informatique? Si oui, quels aspects faut-il prendre en compte? Quels moyens faut-il prendre? Qui doit s'en charger? Nous ne proposons pas d'aborder toutes les questions qui sont pertinentes au thème du colloque, mais plutôt, après un bref survol de la situation générale, de décrire celle, plus précise, de notre enseignement à Laval, pour en tirer quelques leçons.

1.1 *Les inconvénients*

Au départ, il y a une réalité qui s'impose de façon brutale : le français comporte des inconvénients. Vouloir fonctionner en français entraîne certaines conséquences.

1.1.1 L'inconvénient des accents

L'exemple le plus évident est celui des accents. Il vaut la peine d'en parler encore; d'abord parce que les problèmes causés par les accents ne sont pas encore tous résolus, mais aussi parce que l'évolution de la situation nous renseigne sur les attitudes des gens.

Il y a vingt-cinq ans, les claviers les plus courants offraient une gamme réduite de caractères : on ne disposait que de 26 caractères alphabétiques, tous en majuscules, sans accents. A cette époque, il était courant de conclure que le français était mal adapté à l'informatique. C'est seulement plus tard que l'équipement et des logiciels plus évolués sont devenus plus familiers, que ce premier préjugé a commencé à céder à l'idée que c'était plutôt l'informatique qui n'était pas bien adaptée au français.

---

[1]Communication présentée lors du 56e Congrès de l'ACFAS, Moncton, 9-13 mai 1988.

Un événement anodin permet d'illustrer à la fois le progrès accompli et les obstacles qui restent à franchir. L'année dernière, lors d'une conférence internationale sur les bases de données dans les sciences humaines et sociales, un conférencier américain proposait des normes pour la transmission des données à travers les réseaux de télécommunication. Dans l'auditoire, il y avait beaucoup de personnes qui s'occupaient de langues autres que l'anglais. La première recommandation était à l'effet que tout le monde devait renoncer à utiliser des caractères au-delà du code ASCII 128. Le tollé qui a suivi a laissé le conférencier bouche bée.

La transmission des données n'est pas le seul problème relatif aux accents. Les logiciels courants font encore des tris alphabétiques où les lettres accentuées traînent comme des moutons noirs derrière le troupeau des lettres «normales».

TABLEAU 1

L'ordre alphabétique inhérent au code ASCII

a
•
•
cela
•
•
encre
•
•
sut
sèbre
•
•
école
•
•
ça
•
•
être
•
•
ère

Lorsqu'on traite des textes sans distinction des mots quant à la présence d'une majuscule sur la lettre initiale, on se retrouve avec des erreurs d'orthographe. Ce genre de problème réclame une solution.

TABLEAU 2

Accents et majuscules

texte imprimé :

A lui, il a donné un livre.
À lui, il a donné un livre.
... à lui, il a donné un livre.

clavier IBM/PC :

A ... ... a ...
... a ... ... a ...

conversion minuscule :

a ... ... a ...
(forme (forme
de : a) de : avoir)

1.1.2 Le français ne peut pas bien s'adapter

On peut soupçonner que mes qualités d'anglophone et de linguiste portent à croire que je peux parler d'autorité lorsqu'il est question du langage dans le domaine de l'informatique. De nombreuses personnes, rompues ou non à l'informatique, ont abordé avec moi des sujets linguistiques, et les discussions ont souvent dépassé les limites des connaissances bien établies. Parfois, même si les opinions ne sont pas bien appuyées sur le plan scientifique, elles sont révélatrices de malaises ou de préoccupations réelles. Par exemple, une plainte formulée par certains informaticiens francophones est que le français ne peut pas s'adapter convenablement à de nouvelles réalités, notamment celle de l'informatique. Ici, il faut entendre l'évolution du système linguistique, et particulièrement le processus de formation des mots. En caricature, on présente le français comme un objet de grand art, qui, un peu comme les bijoux de la Couronne, serait gardé jalousement par les membres de l'Académie française. Le commun des mortels n'aurait pas le droit d'y toucher.

Il y a une fausse conception de la langue à la base de cette attitude. Les personnes qui partagent cette vision des choses la contredisent dans leurs propres agissements quotidiens, en utilisant allègrement des néologismes, des emprunts et des expressions qui s'écartent de la norme. Mais ils ont le sentiment que ce qu'ils font ne concerne pas le français officiel qui lui, ne court par les rues, du moins à Québec. Ils ignorent que c'est la somme des agissements individuels qui gouverne l'évolution de la langue.

Cette fausse conception est révélatrice. Ce qui est perçu comme la norme l'est en même temps comme inaccessible, ce qui enlève tout l'avantage utilitaire à l'existence de la norme. Une action à prendre pour changer de telles attitudes, si elle est concevable, est sûrement complexe, et devrait impliquer le système d'éducation, ainsi que tous les autres acteurs en aménagement linguistique sur une longue période.

Plus concrètement, ce qui était en cause chez ces informaticiens anonymes était la possibilité de former de nouveaux mots. Prenons un exemple dans le domaine des loisirs, où il a fallu nommer les concepts liés à ce qu'on appelle en anglais *wind surfing*.

TABLEAU 3

La formation des mots

catégorie	anglais	français
le sport	wind surfing	la planche à voile
l'équipement	a wind surfer	une planche à voile
l'action	to wind surf	faire de la planche à voile
l'acteur	a wind surfer	un véliplanchiste

1.1.3 La mentalité nord-américaine

En poussant un peu plus loin dans cette direction, certains francophones en viennent à rejeter le français en faveur de l'anglais en tant que véhicule de l'informatique. Les raisons sont nombreuses, mais il y a des thèmes récurrents, dont le principal est qu'il y aurait une mentalité liée à la langue française qui n'est pas propice à la mentalité de l'informatique. En jargon psycholinguistique, c'est *l'hypothèse Sapir-Whorf*. Une variante contradictoire de cette position veut que les francophones canadiens aient une mentalité nord-américaine, en dépit du fait qu'ils parlent français. Entre ces deux positions, où se situe la vérité?

Ces attitudes, fondées ou non, révèlent un malaise que l'on ne peut pas écarter du revers de la main. Qu'y a-t-il là-dessous? Est-ce la divergence socio-culturelle avec la France et la convergence vers le creuset nord-américain?

Les indices glanés dans certaines conversations permettent de croire à des causes moins globales et moins irréversibles pour ces attitudes. Il y a d'abord les inconvénients mineurs abordés ci-dessus, comme l'exemple des accents. Plus sérieusement, il y a la question de la formation des mots qui vient d'être évoquée, qui fait partie d'un problème plus général de terminologie et de vocabulaire, que les autres conférenciers pourront développer davantage. Il est clair que l'absence de vocabulaire commun est une entrave à la communication, et l'acquisition de ce vocabulaire par l'individu représente un investissement intellectuel considérable. On peut comprendre que ceux et celles qui ont investi du côté de l'anglais en raison du contexte, contribuent maintenant à la résistance à la francisation.

1.1.4 L'anglais serait plus précis

Il reste un facteur que nous voudrions présenter sous la forme d'hypothèse : aux néologismes d'origine française, on préfère les emprunts à l'anglais. La raison invoquée est qu'un terme emprunté à l'anglais pour un concept nouveau est univoque; il signifie par définition exactement ce que l'utilisateur entend. Bien entendu, un mot français peut également par définition signifier précisément la même chose. L'ennui, c'est qu'il peut tout aussi bien référer à plusieurs autres choses. Ainsi, la personne qui parle peut avoir le sentiment de cerner de plus près son idée en utilisant un jargon qui pourrait à la limite venir du swahili. *À la limite*, car la relative familiarité de l'anglais contribue sûrement au phénomène. Souvent on entend dire, de la part d'étudiants ou étudiantes, au sujet des codes de commande pour un langage de programmation, que les termes anglais sont beaucoup plus précis que les équivalents français, par exemple : *print* par contraste avec *imprimer* pour l'affichage à l'écran en **BASIC**. L'ironie, c'est évidemment que pour les anglophones le choix de *print* possède tous les inconvénients de *imprimer* pour les francophones. Les anglophones finiraient sans doute par trouver des qualités intrinsèques au mot *imprimer* qui manquent au mot anglais.

Supposons que cette hypothèse soit fondée. Alors, ne vaudrait-il pas mieux tolérer, ou même encourager l'adoption d'un sous-ensemble de l'anglais pour l'informatique? Ne serait-ce pas analogue à l'adoption d'une notation mathématique, comme pour le calcul différentiel? Dans ce dernier cas, la réponse est clairement non, parce que la notation mathématique constitue un tout cohérent bien circonscrit, relativement autonome et indépendant de la langue. C'est également le cas de chaque langage de programmation, mais ce n'est pas le cas pour la langue de l'informatique en général. Il faut cimenter le jargon avec du mortier de langage ordinaire afin de constituer des phrases. L'effet le plus nuisible de cette tendance à créer des jargons est l'obligation qui est ainsi créée pour les utilisateurs d'acquérir un code supplémentaire pour accéder à l'autel des initiés en informatique.

1.1.5 La francisation des logiciels

Un aspect crucial de la francisation concerne l'adaptation en français des logiciels courants. Certains produits, par exemple **WORD PERFECT**, ont déjà des version françaises très convenables. Cela signifie que la documentation externe, la documentation interne (c'est-à-dire les fonctions d'aide) et l'affichage des messages et commandes sont tous en français. Il faut souligner qu'il ne suffit pas d'un vernis français sur un meuble anglais. Si la touche «p» signifie *print* en anglais, il faut remanier non seulement les messages, mais l'ensemble des protocoles d'échange avec l'utilisateur en français.

Cet état de chose entraîne d'autres effets. Souvent le fait de travailler avec une version française veut dire renoncer à utiliser la version la plus récente. Dans certains contextes, le retard peut entraîner une perte sur le plan de la compétitivité.

La francisation d'un logiciel n'est pas simple sur le plan informatique, ni sur le plan linguistique. Dernièrement, dans le cadre du projet **RELAI**, parrainé par une entente entre l'université Laval et **IBM**, les professeurs Lionel Meney, Denis Juhel et moi-même avons décortiqué le manuel et le guide de l'utilisateur du système d'exploitation **DOS 3.2**. Nous y avons trouvé de nombreux exemples d'emploi de mots qui contribuent à rendre le texte difficile pour un lecteur non averti.

TABLEAU 4

Le vocabulaire du IBM DOS 3.2

sauvegarde
unité, unité de disquette
(lecteur ou disquette)
formater
restaurer
éditer
ESC
ERASE
TYPE
DIR
TREE
ONLINE
CHKDSK
MKDIR

1.1.6 Les publications en français

Il y a également le problème des publications; ou plutôt *les* problèmes. Ceux et celles qui se réclament d'une mentalité nord-américaine citent la difficulté qu'ils ont à lire les publications françaises. Il semble difficile de les contredire. Est-ce simplement un manque de familiarité avec la terminologie française? Ou y a-t-il en France une tradition d'abstraction à l'extrême, de théorisation à outrance et d'obscurcissement savant qui contraste avec celle plus concrète, empirique et directe du monde anglo-saxon? Il semble osé d'affirmer une généralité aussi grossière. Quoi qu'il en soit, il y a une insatisfaction manifeste à l'égard des publications françaises, que ce soit dû au contenu ou au style. Si l'on ajoute à cela le nombre relativement limité de titres, le tableau est plutôt sombre.

2. L'expérience d'enseignement à l'université Laval

Certains arguments plus ou moins sérieux ont été mis de l'avant quant aux inconvénients entraînés par l'utilisation du français. Avant d'aborder le revers de la médaille, et dans le but de justifier ces propos, nous décrirons certaines facettes de notre expérience d'une vingtaine d'années d'enseignement à l'université Laval.

2.1 *Le contexte*

Il s'agit de l'initiation à l'informatique d'étudiants et étudiantes en Lettres. Au départ cela signifie des personnes qui n'ont pas le profil «sciences» du niveau collégial. Une bonne proportion de cette clientèle a une sainte frousse des systèmes formels comme ceux des mathématiques. Peu d'entre elles ont une connaissance préalable de l'informatique.

Le cours en cause s'intitule *Automatisation linguistique*, mis sur pied avant même l'apparition des micro-ordinateurs. L'objectif principal n'était pas de maîtriser un langage de programmation; il y avait à cet effet des cours au département d'informatique. Il s'agissait plutôt d'acquérir les notions

et principes qui sous-tendent l'automatisation dans le domaine de la linguistique. Essentiellement, on voyait les algorithmes qui permettent des manipulations simples de textes en vue de la production d'index et de concordances.

Ceux qui insistaient pour apprendre un langage de programmation pouvaient s'inscrire à des cours d'APL, de FORTRAN ou de PL/1 au département d'informatique. Il devait même y avoir des sections de cours destinées à la clientèle non-scientifique. L'expérience était assez concluante, de façon négative : les étudiants et les étudiantes de la Faculté des Lettres trébuchaient sur des exercices qui leur étaient difficilement compréhensibles, comme faire un programme pour calculer le poids que peut supporter un pont avec une telle structure.

2.2 *Les débuts*

Au début, tout se faisait sans ordianteur, jusqu'à la simulation manuelle du fonctionnement des algorithmes. Voici l'intérêt de ce propos pour le thème du colloque : les algorithmes étaient formulés en langage maison, une sorte de version en français courant d'un langage de programmation non-identifié.

TABLEAU 5

Un programme en langage «maison»

A, B -----> addition -----> C (= A + B)

1. assigner une valeur à A
2. assigner une valeur à B
3. soustraire 1 de A
4. additionner 1 à C
5. vérifier si A est égal à séro
   si oui, aller à 6
   si non, aller à 3
6. soustraire 1 de B
7. additionner 1 à C
8. vérifier si B est égal à séro
   si oui, aller à 9
   si non, aller à 6
9. afficher la valeur de C
10. arrêter

Sans prétendre qu'il répondait à toutes les attentes, le langage «maison» fonctionnait. Les évaluations de cours faites par les étudiants montraient un taux de satisfaction qui m'étonne encore. Heureusement qu'on ne mesurait pas le taux de frustration dû au fait qu'on ne disposait pas d'ordinateurs.

2.3 *Le cours avec micro-ordinateur*

Avec l'introduction des micro-ordinateurs, le cours a évolué, impliquant l'apprentissage du BASIC et l'obstacle du jargon en anglais. La solution a été de continuer à décrire les algorithmes en français en deux étapes dans les premières semaines de cours : une étape pour présenter l'algorithme dans ses grandes lignes en vue d'une compréhension globale, suivie d'une représentation plus détaillée,

suivant les lignes de l'éventuel programme en BASIC. Ensuite, on présentait le jargon précis pour les instructions BASIC. Au besoin, le code pour la commande était justifié par les termes anglais, par exemple *LPRINT* vient de *line printer* et fait imprimer ligne par ligne.

Ainsi, nous avons combiné une présentation en français avec le code autonome qu'est le langage de programmation. La formule s'est avérée satisfaisante pour les objectifs du cours et n'a pas soulevé de rejet général à cause de l'utilisation de l'anglais pour le langage de programmation. Une minorité ne réussissait pas à avaler la couleuvre, mais il est difficle de faire le partage entre l'anglais et la programmation comme sources de difficulté.

Il y a plus important que le langage de programmation en tant que tel : la langue utilisée dans la documentation. Les étudiants résistent à lire, même en français, et nous en avons des preuves convaincantes : les fiches d'emprunt des volumes déposés à la bibliothèque pour le cours sont presque vierges. En ce qui concerne la lecture en anglais, certains étudiants et étudiantes admettent franchement leur inaptitude. Quant aux manuels et guides, il s'agit d'un genre littéraire assez spécial, et le problème est surcomposé.

Évidemment, l'idéal serait de tout avoir en français : langage de programmation, système d'exploitation, guides, manuels et lectures d'appoint. Pour un cours d'initiation, c'est maintenant une possibilité. Le choix de langage est plus limité, mais il suffit d'un seul dans la gamme appropriée : TURBO-PASCAL pourrait remplacer le BASIC.

3. Les avantages du français

Revenons maintenant au sujet des avantages en faveur du français dans le monde de l'informatique. La position adoptée sera claire. Un certain nombre de désavantages ont été évoqués ci-dessus, qui sont plus ou moins difficiles à évaluer. Il y a une contrepartie, dont la portée exacte est également difficile à apprécier. Il peut ne pas exister de solution générale, valable pour toutes les situations : qu'on pense aux ordinateurs centraux par opposition aux micros, aux divers niveaux scolaires, à la division entre informaticiens professionnels et utilisateurs de l'informatique. Les choix sont ultimement des choix de société et comme tels impliquent des décisions politiques. De telles décisions doivent être prises en considérant les coûts et les bénéfices, aussi difficile que ça puisse être.

Le débat autour de la francisation doit certainement faire une large place à des considérations d'ordre social. Les besoins de communication à l'intérieur d'une société sont évidents. La volonté de cohésion peut entraîner celle de la préservation d'un rôle prépondérant pour la langue dans tous les domaines d'activité. Sans s'éterniser là-dessus, tirons plutôt un argument d'une expérience d'initiation des étudiants et étudiantes à l'informatique en linguistique. Cette expérience conduit à un constat qui constitue un argument prévalant sur tous les autres dans des contextes comparables. Ce constat est le suivant : certains étudiants et étudiantes francophones ont moins de difficulté à s'initier à l'informatique en français. L'exigence du recours à l'anglais constitue un obstacle pour une partie de la clientèle. Si cela est vrai des étudiants et étudiantes universitaires en Lettres, ça l'est d'autant plus pour la population en général.

Cette conclusion est d'une évidence telle, que l'on serait en droit de croire que nous perdons notre temps à la proclamer. Mais elle a des corollaires un peu moins évidents. L'aspect le plus important pour l'avenir de la société est celui de l'accessibilité de l'informatique, qui se doit d'être la plus large possible. Toute limitation de l'accès tendra à nuire au développement économique, tout en exacerbant les inégalités sociales. Si le fait de promouvoir l'informatique en français permet en même temps de renforcer la position de la langue française, tant mieux.

L'INFORMATIQUE EN FRANÇAIS[1]

Conrad Ouellon
Langues et linguistique
Université Laval

Introduction

La question du français dans l'informatique est apparentée au problème du français langue de travail, mais elle le déborde puisque l'informatique s'est maintenant infiltrée dans bien d'autres secteurs de notre activité.

L'innovation technologique, règle générale, est en soi image de la société dans laquelle elle s'est d'abord implantée. Elle reflète donc à des degrés divers les valeurs culturelles de la société dont elle est issue. L'importation de nouvelles technologies, dans des contextes particuliers, favorise certains transferts culturels; de façon plus particulière, elle est aussi susceptible de modifier les comportements linguistiques de la société importatrice, la langue se trouvant être l'un des modes privilégiés d'expression de la culture.

Si la nouvelle technologie ne déborde pas les limites d'un laboratoire ou d'une usine, les conséquences du changement ne prêtent peut-être pas à conséquence. Il en est tout autrement si on considère le cas de l'informatique. La société qui s'informatise tend, en effet, à le faire à tous les niveaux ou presque, et dans toutes les sphères de son activité.

La société québécoise, de par sa position géographique, est une enclave dans un monde anglophone et les technologies de l'information lui viennent principalement des États-Unis. En conséquence, l'initiation à l'informatique s'est d'abord faite en anglais, à travers les premiers manuels et guides d'utilisateur; les informaticiens et autres spécialistes de l'informatique travaillaient avec une terminologie anglaise. Cette terminologie était aussi omniprésente sur les pièces d'équipement, les touches de clavier, dans les manuels et, évidemment, dans les langages de programmation ...

Cependant, à cause justement de la généralisation de l'informatisation, il se dessine de plus en plus une volonté nette de la rendre accessible en français. Sans nous attacher aux raisons d'ordre politique qui justifient un effort plus grand de francisation de l'informatique, nous nous placerons plutôt du point de vue des avantages que la société québécoise peut en retirer sur les plans de la productivité et de la qualité de l'expertise qu'il est possible de développer au Québec. Pour plusieurs, ces arguments valent plus que les autres et on peut le comprendre. Cependant, tout devient politique en conclusion.

État de la question

Il semble important dès le départ de faire la distinction entre la langue en usage dans les gros systèmes informatiques et celle qu'utilise le monde de la micro-informatique. Que les gros systèmes

[1]Conférence prononcée devant les membres du Mouvement national des Québécois, Québec, avril 1988.

invisibles à tous, sauf aux analystes et aux programmeurs, fonctionnent avec du logiciel anglais et du matériel de support dans la même langue ne dérange guère pour l'instant, compte tenu de la situation d'urgence qui prevaut ailleurs. Toutefois, les micro-ordinateurs, peu importe leur usage, devraient communiquer en français avec les personnes qui les utilisent et tout le support devrait être fourni dans cette langue. Le même principe devrait prévaloir pour les systèmes de communication entre les micro-ordinateurs et les systèmes centraux. Bref, une technologie qui se veut à l'avant-garde et de plus en plus à l'écoute de l'utilisateur devrait se plier à un premier grand principe qui veut qu'un travailleur ait le droit de travailler dans sa langue maternelle, ici le français. Il ne paraît pas non plus insensé de penser que c'est à la machine à s'adapter à son utilisateur et non l'inverse. C'est une question d'ergonomie, tout aussi importante que les précautions apportées à la conception de l'ameublement et à la qualité des écrans par exemple.

On peut penser, sans faire d'enquête, qu'un utilisateur d'informatique qui emploie sa langue maternelle mettra moins de temps à se former, sera moins sujet à des problèmes de compréhension des messages et sera par conséquent plus productif. A la condition évidemment qu'il ne soit pas obligé de recourir à la version anglaise pour comprendre les messages qu'il reçoit ou les manuels qu'il consulte. Cette question sera abordée plus loin. Il faut aussi signaler que le problème n'est pas le même selon qu'on parle de l'utilisation du français pour quelqu'un qui s'initie à l'informatique en français ou pour quelqu'un qui doit migrer d'un environnement linguistique anglais à un autre en français.

À l'heure actuelle, la situation semble varier considérablement. Voici, à titre d'illustration, quelques exemples qui font clairement ressortir qu'il est possible de travailler en français avec un ordinateur.

a) Une compagnie d'assurance d'origine américaine comme **La Métropolitaine** a tenu, et réussi, à fournir un environnement informatique entièrement francophone à ses employés québécois. Qu'en résulte-t-il? Un haut taux de satisfaction des employés à peu près tous unilingues français et une productivité accrue dans une succursale récemment ouverte dans une région du sud de Québec, bien connue pour son dynamisme.
b) Un autre exemple est celui de la **Banque d'épargne** de la Cité et du district de Montréal, qui a terminé la francisation de son système informatique. On note dans un rapport déposé en octobre 1986 le taux élevé de satisfaction des usagers, mais on y souligne aussi le coût à payer pour l'institution.
c) Plusieurs systèmes informatiques en usage dans les services gouvernementaux, dans les centres hospitaliers, fonctionnent également en français.

Les exemples de ce type de réussite sont nombreux. Il semble bien que les sociétés qui le veulent possèdent un service informatique en français. Elles en retirent une meillleure image et font volontiers état, avec raison, de leur respect envers leurs travailleurs et leur clientèle francophones; elles font à peu près toutes valoir un fort taux de satisfaction de leurs employés.

À l'appui de ce qui précède, un rapport du **Comité Canada-Québec** (formé par les ministères des Communications des deux paliers de gouvernement), sur le développement du logiciel d'expression française (1986), signale aussi les inconvénients à travailler dans une autre langue : «Des études ont démontré que les logiciels conçus en anglais désavantagent les utilisateurs francophones dont l'accès aux techniques nouvelles se trouve ainsi freiné. Le plus souvent américains, ces logiciels sont mal adaptés à la réalité des groupes francophones. Il en résulte une perte d'identité culturelle et de maîtrise linguistique». (*Rapport du Comité Canada-Québec sur le développement du logiciel d'expression française* 1987:3).

Il faut toutefois reconnaître que les PME, et évidemment les particuliers, qui n'ont pas les moyens d'investir dans la francisation, éprouvent des difficultés de divers ordres :

a) impossibilité d'acquérir les versions récentes des logiciels d'usage commun en français (ex. Word **Perfect 5.0** disponible en anglais, pour la version française un an plus tard);
b) difficulté sinon impossibilité d'acquérir plusieurs logiciels, qui ne sont pas distribués sur le marché francophone;
c) agacement devant les traductions boiteuses et même devant les traductions en français hexagonal, ce qui pose la question de la norme du français;
d) difficultés de compréhension des manuels d'utilisation et des guides d'entretien en anglais.

La conséquence est le recours au matériel anglais pour ceux qui connaissent suffisamment cette langue, les francophones unilingues devant se contenter d'un matériel désuet dans plusieurs cas.

Quelles sont les réactions possibles de l'utilisateur devant ces faits? Elles sont diversifiées. Un professeur d'informatique manifestait récemment son indifférence totale devant ces problèmes. Forte réticence également de certains ministères du Gouvernement du Québec qui soutiennent qu'on ne doit pas sacrifier la rentabilité au français. La réticence est encore plus forte chez ceux qui sont déjà habitués à travailler en anglais; nous ne citerons qu'un seul exemple, celui des services informatiques d'une certaine université qui distribue encore des manuels anglais pour la formation à certains logiciels alors que la documentation française est disponible (ils le font à la demande, bien souvent, des usagers).

Plusieurs utilisateurs réagissent autrement et exigent l'informatisation en français. La demande est de plus en plus forte et c'est en tant que consommateurs que plusieurs demandent à être servis dans leur langue ...

Existe-il des perspectives d'amélioration? Évidemment, il en existe, et nous nous y attarderons maintenant. Dans le contexte de concurrence féroce qui s'est développé dans le monde de l'informatique, c'est le client qui a raison, qui aura ce qu'il exige. Un représentant de la **Société IBM** l'affirmait sans aucune hésitation à un colloque en février 1988. (Il disait aussi que sa société avait consulté récemment un groupe d'experts francophones (région de Montréal) sur l'intérêt qu'il y aurait à traduire plusieurs logiciels très populaires en Amérique de Nord; on a répondu que ce serait une perte de temps et qu'on pouvait très bien se contenter du produit anglais ...). Une chose est certaine, les grandes sociétés sont plus sensibles qu'on le croit souvent à la pression des consommateurs et à leur image de bon citoyen; et surtout, elles sont bien placées pour forcer les producteurs de logiciels comme **Microsoft**, **Ashton-Tate**, etc., à tenir compte des exigences des consommateurs.

Problèmes rencontrés par les utilisateurs francophones

De façon concrète, quels sont les problèmes que rencontre l'utilisateur francophone, ou plus généralement non-anglophone, en informatique? En voici quelques-uns, sans prétendre qu'ils soient tous de la même importance. On peut s'entendre pour dire qu'ils relèvent de trois aspects de l'informatique : l'équipement, les logiciels, les manuels.

1- *Les claviers*: il y a bien sûr la multiplicité des claviers. On peut dire que le problème est le même pour l'ensemble des utilisateurs. Il n'est cependant pas du même ordre pour tous. L'anglophone est moins touché puisque les 10 chiffres, les 26 lettres et, dans une moindre mesure les signes de ponctuation, occupent toujours la même place sur le clavier. Le francophone, lui, est aux prises avec la question des accents très fréquents en français, des accents qui doivent apparaître sur les majuscules aussi (*Règles typographiques des éditeurs français* 1975). Ces accents sont tous produits, quand on le peut, par des frappes doubles. De plus, ils n'occupent pas toujours le même emplacement sur les claviers, forçant même à l'occasion le déplacement de caractères ailleurs sur le clavier, d'où l'embarras pour les utilisateurs.

Inutile de mentionner que la situation est la même avec les machines à écrire conventionnelles. De plus, une sorte d'anarchie dans la conception du matériel informatique oblige à configurer imprimantes et écrans pour que les caractères soient vraiment accentués sur ces deux types de périphériques. Ce pourrait être un mal encore acceptable si le fait de ne pas reconnaître les accents comme de véritables caractères en français ne créait pas des problèmes encore plus sérieux dans le tri de documents et la recherche d'information dans les textes. Par exemple, *du* et *dû* sont considérés comme un même mot du point de vue de la langue anglaise. Un programme de recherche qui ne tient pas compte de l'accent devient difficilement utilisable pour un francophone. Autre difficulté pour les caractères supportés par les claviers : la distinction entre les ligatures *ae* et *oe* et les diphtongues *ae* et *oe* ne peut se faire (*boeuf* et *coefficient* ...).

Enfin, on identifie le plus souvent les touches en langue anglaise: *delete*, *escape*, *insert*, *scroll*, *break* font maintenant presque partie de la langue courante.

2- *Les logiciels.* Ils sont souvent en anglais, à moins qu'on appartienne à une grande organisation qui a les moyens et la volonté de les franciser. Souvent, ces logiciels ne sont pas en mesure de faire convenablement des opérations de tri ou de classement alphabétique en français (la version assez récente de **DBase-111+** ne le pouvait pas).

Les analyseurs de texte, les lemmatiseurs sont souvent conçus pour l'anglais et ne peuvent guère rendre service en français.

Les lecteurs optiques en mesure de décoder les accents sont encore peu nombreux et coûtent plus chers.

3- *Les manuels.* Ils sont souvent mal traduits. Les traductions faites en France agacent aussi les locuteurs francophones nord-américains. Les principaux problèmes rencontrés sont la polysémie, les ambiguïtés, etc. Il se fait quand même de plus en plus d'excellentes traductions. Une société comme **IBM** investit suffisamment dans ses services montréalais de traduction et de terminologie pour obtenir des produits intéressants.

Solutions possibles

Existe-t-il des solutions aux problèmes qu'éprouvent les utilisateurs francophones? La réponse est affirmative, on ne peut guère en douter.

a) En ce qui concerne l'équipement, on constate qu'il s'établit certains standards. Les pas qui restent à faire pour qu'on s'entende sur un modèle uniforme de clavier, du moins en Amérique du Nord, sont moins grands qu'on pourrait le croire. Il en est tout autrement des codes ASCII et EBCDIC qui sont probablement là encore pour longtemps, mais qui ont la caractéristique de limiter le nombre de caractères disponibles. Il n'y a pas de grands obstacles à ce que les accents soient considérés comme de véritables caractères pour des fins de tri et de recherche. Plusieurs compagnies le font.

La francisation des touches constitue peut-être un faux problème; il suffirait probablement qu'il y ait concertation sur un choix d'icônes pour représenter les diverses fonctions. On a bien réussi à s'entendre sur une représentation «internationale» de la signalisation routière! Une forte pression de la clientèle devrait permettre qu'on finisse par avoir raison sur ces points. Certains claviers récents affichent d'ailleurs leurs fonctions en langue française.

b) Le problème que posent les logiciels est d'un autre ordre. C'est probablement dans la modularisation que réside la solution. Au lieu de s'attacher à produire des logiciels unilingues, il faudrait peut-être carrément aller vers la conception de logiciels multilingues. Un logiciel multilingue serait constitué d'une base de programmation en français par exemple, alors que toutes les autres fonctions imaginables, les messages à l'écran, les fonctions de tri, de recherche, les données spéciales comme la façon d'indiquer la date, l'heure, les systèmes de mesure, etc., pourraient constituer la matière de plusieurs modules interchangeables faits en fonction de clientèles de culture et de langue différentes. Des logiciels ainsi construits, et il en existe déjà paraît-il, feraient sûrement une forte concurrence aux logiciels unilingues. C'est là un des champs où nous pouvons transformer en avantage notre position de minoritaires du point de vue linguistique. Il ne serait en effet pas toujours avantageux de construire des logiciels en langue française au Québec. Le marché est restreint et il ne faut pas compter sur la France pour rentabiliser le produit. D'un autre point de vue, nous sommes déjà habitués (trop diraient certains) à tenir compte de la présence de gens parlant d'autres langues. On voit facilement le marché qui s'ouvre au logiciel multilingue. En effet, il n'y a pas que le Québec qui subisse la pression anglicisante de l'informatique. C'est vrai de tous les pays où l'anglais n'est pas la langue officielle. C'est même vrai de la Grande-Bretagne qui n'accepte pas toujours l'usage américain.

c) Enfin, il reste la question des manuels et guides en langue anglaise. Il faut les exiger en langue française et on peut souvent les obtenir dans cette langue. On est cependant en droit de s'interroger sur la qualité de cette traduction, bien qu'il faille reconnaître sous ce rapport les efforts des grandes sociétés. Les services linguistiques montréalais de la **Société IBM** fournissent des traductions ou adaptations en excellent français. Signalons cependant qu'elle ne partage pas nécessairement ses travaux avec la **Société IBM-France**, reconnaissant ainsi certaines divergences entre les façons française et québécoise de s'exprimer, ce qui n'est pas surprenant, puisque les cultures sont différentes. L'une des façons possibles de régler ce problème de traduction, serait peut-être de faire produire ces manuels en français au départ plutôt que les traduire. La traduction entraîne aussi des frais énormes et des retards dans la mise en circulation des logiciels; on parle de retards de l'ordre de neuf mois ...

On le voit aisément, il n'existe pas de solution facile à tous les problèmes qui ont été soulevés. Les véritables solutions relèvent d'un effort soutenu de la recherche linguistique.

Recherches à envisager

Il faudra résolument s'attacher à définir une forme de français écrit et parlé en vue de son usage par l'ordinateur. D'une certaine façon, comme le disait Richard Parent (ministère des Communications) dans une allocution prononcée en mai 1986, «la langue française a besoin d'être reconstruite dans l'espace électronique à divers niveaux (grammaires, réseaux sémantiques, règles de dialogue, etc.)» (Parent 1986:6).

Il est de la plus haute importance, pour les usagers francophones de l'informatique, que leur langue puisse être aussi bien traitée par l'ordinateur que l'anglais, d'où la nécessité de disposer de descriptions de la langue réalisées à l'intention des ordinateurs, comme les dictionnaires ont été faits pour les humains.

En bref, il faut en arriver à définir une version formalisée du français dont les rapports avec le français habituel seraient du même ordre que ceux qui existent entre l'anglais informatique et l'anglais courant. Ce sont des langues techniques dont le seul mérite est de faire obéir une machine. Une telle forme de langue pourrait probablement se prêter à la traduction automatique entre autres avantages ... Si l'on considère que l'informatique envahit constamment à peu près toutes les sphères

d'activités humaines, on mesure l'ampleur de ce défi. Et les participants au *Sommet des pays francophones* de Québec (septembre 1987) l'ont bien compris en proposant de favoriser les industries de la langue. Restent à voir les retombées réelles des efforts annoncés, si jamais on parvient à se dégager des contraintes d'ordre politique. Il faudra quand même travailler en collaboration avec les autres pays francophones, surtout la France, dans cette grande entreprise de «mise à plat» de la langue française. Plusieurs chercheurs et de nombreuses entreprises s'y intéressent de plus en plus. Au Québec même, des centres de recherche comme le **CRIM** (Centre de recherche en informatique de Montréal) et le **CCRIT** (Centre canadien de recherche en informatisation du travail) à Montréal, le **CEFRIO** (Centre francophone de recherche en informatisation des organisations) et le **CIRB** (Centre international de recherche sur le bilinguisme) à Québec, y travaillent en collaboration avec les services gouvernementaux et les industries.

L'informatique en français? Oui, sans aucune hésitation. C'est le droit légitime de travailler dans sa langue d'usage qui est en jeu. Du point de vue des organisations, le recours à la langue maternelle en informatique est sans aucun doute intéressant du point de vue de la productivité. Il est sûr qu'il y a des coûts pour franciser l'informatique. Cependant, si l'opération est menée de façon méthodique et de concert entre tous les intervenants, de disciplines diverses et d'organisations variées, ce qui semble un handicap au premier abord peut permettre de bâtir un secteur d'excellence, permettre par conséquent de développer une expertise «exportable».

Enfin, il y va peut-être de la survie même de la langue française de prendre tous les moyens pour qu'elle serve d'outil de communication entre les personnes, mais aussi entre les personnes et la machine, et peut-être aussi de machine à machine ... Comme le disait Bernard Cassen, chef pour la France de la **Mission Industries de la langue** : «On peut ... dire que les langues qui ne s'industrialisent pas cesseront à un terme plus ou moins bref, d'être véhiculaires, parfois même, dans d'importants secteurs d'activités des pays où elles sont parlées» (Cassen 1986:13).

BIBLIOGRAPHIE

CASSEN (Bernard), Un nouveau front pour le français et la langue de l'Europe, *Les industries de la langue. Enjeux pour l'Europe. Actes du colloque de Tours.*, Tours 28 février-1er mars 1986, dans *Encrages*, n° 16, novembre 1986, p. 13.

PARENT (Richard), *Le français en informatique à l'heure des systèmes experts*, Allocution prononcée devant les membres de l'Association de la recherche en communication du Québec, Montréal, 8-9 mai 1986, Ministère des Communications du Québec, p. 6.

Rapport d'étape du Comité Canada-Québec sur le développement du logiciel d'expression française, Sommaire et faits saillants, Gouvernement du Canada - Gouvernement du Québec, juin 1986, p. 3

Règles typographiques des éditeurs français, 1975.

TECHNOLOGIE ET SYNTAGME: LE CAS DE LA MICRO-INFORMATIQUE[1]

Jean Quirion
Langues et linguistique
Université Laval

La présente communication fait état de l'avancement du projet de recherche sur la syntagmatique terminologique en informatique du département de Langues et Linguistique de l'université Laval, qui commence sa deuxième année. L'équipe est composée de Jean-Claude Boulanger et Dorothy Nakos, professeurs responsables, et de deux étudiants au niveau de la maîtrise en terminologie et traduction. La recherche consiste à faire une étude comparative de la structure et du mode de formation des syntagmes en anglais et en français. C'est une tâche de grande envergure qui explore un aspect fondamental de la terminologie. Est-il besoin de rappeler que la proportion de syntagmes dans le vocabulaire des langues de spécialité est généralement évaluée à près de 85%? De par son caractère descriptif, le syntagme est un moyen privilégié de nommer la grande quantité de nouvelles réalités qui apparaissent quotidiennement.

Afin d'éviter toute confusion, signalons la définition du syntagme tel qu'on le connaît en terminologie. Il s'agit d'un groupe de mots d'au moins deux unités linguistiques, syntaxiquement liées, qui fonctionnent comme une unité simple. Ce signifiant ne dénomme qu'un et un seul signifié, dans un domaine donné de la connaissance humaine. Cette définition s'inspire de Jean-Claude Boulanger (Boulanger 1988:22).

Le «syntagme» en terminologie s'éloigne de la notion de syntagme utilisée en linguistique. *Syntagme* n'est que l'un des 25 termes relevés par Rostislav Kocourek en 1979 pour dénommer cette notion. Entre autres, Benveniste parlait de *synapsie*, Pottier de *lexie complexe*, Auger de *mot complexe*, Phal préférait *groupe de mots*, Hollyman, *dénomination complexe* et Rey, *groupe lexicalisé*.

La bibliographie analytique constituait la première étape du projet. C'est un volet plutôt mineur, qui de toute façon a été présenté au dernier colloque IBM-Laval en novembre 1987. Cette bibliographie est presque terminée. Près de deux cents ouvrages de tous genres ont été dépouillés : actes de colloques, articles de périodiques, monographies, etc. Les ouvrages retenus ont fait l'objet d'un article. Chaque article contient une référence complète, des descripteurs et un résumé de l'ouvrage. Les 75 articles retenus en sont au stade de la révision et devraient faire l'objet d'une publication à la fin de l'année. Initialement consignée sur le logiciel **BIBELO**, la bibliographie a été convertie sur le logiciel de traitement de texte **Wordperfect** en vue de la publication.

Par ailleurs, un corpus français a été constitué afin de mettre au banc d'essai les théories recensées dans la bibliographie. Il est constitué de trois ouvrages publiés par l'Office de la langue française du Québec. Il s'agit de :

---

[1]Communication présentée lors du 56e Congrès de l'ACFAS, Moncton, 9-13 mai 1988.

- Terminologie de l'informatique, paru en 1983;
- Vocabulaire du logiciel, paru en 1987;
- Vocabulaire du micro-ordinateur, paru en 1986.

Le corpus contient plus de 6 000 termes, emmagasinés sur un support informatique à l'aide du système de gestion de base de données **dBase III+**. La banque de données linguistiques **IBMot**, fournie par les Services linguistiques de la **Société IBM Canada Ltée**, sert de corpus de réserve.

Chacun des 6 000 termes figure dans cinq bases de données, ou fichiers. Chaque fichier répond à un besoin d'analyse particulier soit :

- **les signes diacritiques (tiret, apostrophe, guillemets, barre oblique, etc.);**
- **les prépositions, ou joncteurs (nombre, sens, fréquence de chacun, etc.);**
- **les caractères morphologiques (emprunt, sigle, mot-valise, etc.);**
- **les renseignements généraux (variante orthographique, synonyme, abréviation, etc.);**
- **et, finalement, la typologie.**

C'est celle-ci qui nous intéresse, car le projet visait avant tout l'établissement d'une typologie renouvelée des syntagmes et l'étude de leurs modes de formation.

C'est d'ailleurs un projet qui ne manque pas d'intérêt. La dernière - et la première - typologie des syntagmes a été établie par Louis Guilbert, il y a plus de vingt ans. Depuis, personne ne s'est vraiment penché sur la question.

Il s'agit de réviser et de compléter la typologie de Guilbert. Ce dernier recensait six formes de syntagmes :

1- Nom + adjectif
ex. : ■ menu déroulant
■ souris optique

2- Adjectif + nom
ex. : ■ haute définition

3- Nom + joncteur prépositionnel + nom
ex. : ■ archivage en caractère
■ numériseur à cylindre

4- Nom + joncteur prépositionnel + prédéterminant + nom
ex. : ■ guide de l'opérateur
■ profil de l'utilisateur

5- Nom + absence de joncteur + nom
ex. : ■ carte paramètre
■ fichier article

6- Nom + joncteur prépositionnel + verbe
ex. : ■ machine à calculer

L'état actuel des recherches indique que ce nombre pourrait être augmenté à peut-être une dizaine, douze tout au plus.

La recherche a révélé jusqu'à maintenant l'existence de modèles différents. Par exemple, on relève la présence de pronoms relatifs (*circuit 5 dont 2*), de conjonctions (*régulation proportionnelle, intégrale et dérivée*) et d'adverbes (*récepteur seulement*). Il faudra trancher : ces syntagmes se réduisent-ils aux modèles avancés par Guilbert? S'agit-il de nouveaux types? Ou encore de formes fautives, de termes simplement mal construits (du type *auto-lave*, depuis remplacé par *lave-auto*)?

Chose certaine, l'économie de la typologie sera privilégiée. Pour certaines écoles, un syntagme long de plus de vingt mots pourra représenter un modèle à lui seul (*nom + adj + prép. + nom + prép. + adj + nom*, etc.). Ainsi, *système de gestion de base de données* peut être considéré comme un syntagme du genre *nom + prép. + nom + prép. + nom + prép. + nom*. Cependant, les chercheurs du projet avancent une hypothèse, dite de «neutralisation», qui reste à vérifier. Selon leur conception, le même syntagme serait considéré comme correspondant au modèle *nom + prép. + nom*. En effet, un «système de gestion» pouvant être déterminé par autre chose qu'une «base de données», le syntagme est décomposé comme suit :

FIGURE 1

système de gestion		de		base de données
n + p + n	+	p	+	n + p + n
fonction nominale	+	p	+	fonction nominale
n	+	p	+	n

Cette recherche amène à poser des questions capitales : qu'est-ce qu'une notion ? Un adjectif constitue-t-il une notion à lui seul ? Doit-il figurer à ce titre comme entrée dans un vocabulaire ?

On trouve par exemple dans le corpus l'entrée *à bascule*. On peut penser qu'il s'agit là d'un nouveau modèle, *prép. + nom*, auquel n'avait pas pensé Guilbert. Mais s'il s'agissait de quelque chose d'incomplet en soi, si ce n'était pas une notion? De toute façon, le terme ne sera-t-il pas nécessairement repris dans l'entrée du nom qu'il qualifie ? Si tel est le cas, il deviendrait inutile de le faire figurer seul en entrée, sauf si on le considère comme un modèle.

Par ailleurs, d'intéressants problèmes sont soulevés en ce qui a trait aux prépositions. Par exemple, *DOS* est rendu de trois façons en français :

1- système d'exploitation de disques (prép. de);
2- système d'exploitation à disques (prép. à);
3- système d'exploitation disques (absence de préposition).

En plus de la différence de modèle entre le dernier syntagme et les deux premiers (*nom + absence de joncteur + nom* par rapport à *nom + prép. + nom*), on note une différence complète de sens entre *système d'exploitation à disques* et *système d'exploitation de disques*. En effet, la préposition *de* signifie dans ce contexte «système qui exploite des disques» tandis que le joncteur *à* renvoie à l'idée que le système exploite «à l'aide de disques».

Ce qui précède est un bref aperçu des nombreuses questions qui se posent au début même de l'analyse. Il en existe plusieurs autres. Toutes ne sauront être résolues par l'étude entreprise car un choix s'impose, étant donné l'envergure des recherches. Mais le fait même qu'elles soient posées est en soi un signe de progrès.

Les résultats de cette recherche serviront à poser des jalons pour la création judicieuse et appropriée de syntagmes. La typologie retenue devrait en principe être applicable à tous les domaines de spécialité.

Dans une optique traductionnelle, il existe quatre possibilités de rendre un terme anglais en français :

- **terme simple E** [input]
  terme simple F [intrant]

Ce n'est cependant pas l'objet de notre étude.

- **terme simple E** [indexing]
  terme syntagmatique F [recherche indexée]

- **terme syntagmatique E** [information processing industry]
  terme simple F [informatique]

- **terme syntagmatique E** [initial program loader]
  terme syntagmatique F [procédure de chargement initial].

La typologie sera utile aux «fabricants» de lexiques et de vocabulaires en approfondissant le sens des syntagmes, en précisant le sens des prépositions, leur fréquence d'utilisation, en fournissant des statistiques sur les modes de formation ainsi que sur les combinaisons interlangues, comme le démontre la série d'exemples précédents.

Le projet a des retombées intéressantes qui seront brièvement décrites ci-après.

Tout d'abord, en plus des deux professeurs responsables du projet, six étudiants et étudiantes y travaillent ou y ont travaillé.

La syntagmatique a également été retenue comme thème de deux séminaires de terminologie, à l'automne 1986 et à l'automne 1987. Les étudiants ont été amenés à se prononcer sur diverses facettes du syntagme par le biais d'exposés et de recherches appliquées.

La sensibilisation à cet aspect fondamental de la terminologie a éveillé chez certains le désir d'en poursuivre l'étude. En effet, quelques étudiants inscrits à la maîtrise en terminologie axeront une partie de leur essai sur le traitement des syntagmes dans leurs corpus respectifs.

De plus, de nombreuses questions soulevées par l'étude demeurent sans réponse. Les corpus sont là, disponibles pour des recherches approfondissant l'un ou l'autre des aspects problématiques.

Fait non négligeable, le projet débouche également sur la distinction entre le syntagme, le phraséologisme et la phrase. On passe alors «du terme au texte», pour reprendre une expression utilisée par Monique Cormier lors du Colloque sur l'histoire de la terminologie à Bruxelles en mars 1988.

En fait, on sait bien peu de choses sur le phraséologisme, si ce n'est qu'il s'agit d'un groupe de mots ayant un sens donné, mais ne renvoyant pas nécessairement à une seule notion et qu'il se situe quelque part en amont des collocations, des expressions et des locutions et en aval des termes et des syntagmes. Nous examinerons d'ailleurs la possibilité de faire du phraséologisme le sujet de notre

essai de maîtrise. Nos recherches bénéficieraient alors naturellement de l'expérience acquise avec le projet de recherche sur la syntagmatique en ce qui a trait au classement, à la définition, l'identification, la typologie, etc.

Le phraséologisme se situe à une étape postérieure : on quitte alors le lexique, on se rapproche de la syntaxe. Mais avant de pouvoir le faire, il fallait pousser plus en profondeur les recherches lexicales, ce qui est maintenant en bonne voie. Les étudiants comme les professeurs travaillant au projet ont également eu l'occasion de présenter des communications sur le sujet. Ainsi, plus de huit conférences ont été présentées en moins de neuf mois. La recherche en cours a également fait l'objet de discussions lors de récents colloques européens.

De plus, certains étudiants ont la chance d'effectuer des stages en entreprise, d'y rencontrer les utilisateurs de terminologie et de se frotter à l'exercice pratique qui consiste à résoudre des problèmes d'ordre syntagmatique. Un étudiant a pu contribuer à un aspect essentiel du projet en travaillant pendant une période de trois mois au dictionnaire *Le Robert*, à Paris, à l'automne 1987. Il a abordé le problème du traitement du syntagme dans les ouvrages lexicographiques généraux, notammment le *Grand Robert de la langue française*.

À l'heure actuelle, nous effectuons un stage aux Services linguistiques de la Société IBM à Montréal. Il ne suffit pas de parler de terminologie, il faut encore en faire. En contrepartie, de telles recherches permettent à l'entreprise de bénéficier de retombées tant théoriques (lire résultats) que pratiques (lire formation du personnel).

L'importance de toutes ces retombées mérite d'être soulignée car la relève est essentielle afin que le Québec demeure le précurseur en terminologie.

BIBLIOGRAPHIE

BOULANGER (Jean-Claude), 1988, "Le syntagme en informatique : un projet de recherche", *Terminogramme*, janvier 1988, n° 46, p. 22-23.

LE CENTRE D'ANALYSE DE TEXTES PAR ORDINATEUR (ATO)
Un laboratoire de recherche spécialisé dans l'analyse de textes par ordinateur[1]

Jules Duchastel
Université du Québec à Montréal

1. Un Centre d'analyse de textes par ordinateur (ATO)

1.1 *Bref historique*

La création du **Centre d'ATO** est le couronnement d'une déjà longue expérience de recherche en analyse de textes par ordinateur à l'UQAM. Les chercheurs qui s'y adonnaient étaient conscients de la précarité des ressources et de l'absence de garanties de l'université au sujet de la poursuite de leur recherche. Il fallait inciter celle-ci à reconnaître l'originalité et la valeur de ces travaux dans un domaine de pointe, qui s'étaient spontanément développés dans ses murs. En effet, à partir des années 1970, un petit groupe de professeurs-chercheurs et d'étudiants de deuxième et de troisième cycles avaient entrepris de développer des systèmes informatisés capables de traiter et d'analyser l'information textuelle. Au fil des ans, un nombre croissant de chercheurs se sont tournés vers ces systèmes pour effectuer des recherches dans des disciplines aussi diverses que la philosophie, la sociologie, les communications, la psychologie et la linguistique.

C'est dans ce contexte que l'université du Québec à Montréal a progressivement trouvé les moyens d'assurer une certaine stabilité des ressources humaines, logicielles et matérielles nécessaires à ce nouveau type d'activité de recherche.

1.2 *Trois statuts*

Ainsi, dès octobre 1983, les chercheurs en **ATO** ont été officiellement reconnus comme étant un regroupement de chercheurs, ce qui leur permettait de financer l'embauche des premiers membres d'une équipe d'analystes-programmeurs et d'agents de recherche spécialisés dans la manipulation des environnements informatiques. Le nombre de chercheurs n'allait cesser d'augmenter jusqu'à ce jour : l'équipe comprend actuellement huit personnes à temps plein. Cette équipe constitue la première réalisation du Centre. Les personnes qui la composent sont directement confrontées aux problèmes du développement informatique nécessaire à tout projet, qu'il émerge de l'équipe même ou des professeurs-chercheurs rattachés au Centre. L'équipe s'emploie également à la recherche et au développement des environnements de programmation dans leur aspect général, afin qu'ils soient utiles à tout genre d'applications futures.

La seconde reconnaissance officielle est venue du Fonds pour la formation des chercheurs et l'aide à la recherche (FCAR), organisme rattaché au ministère de l'Éducation du Québec, et qui offre des subventions. En raison du caractère unique des activités du Centre, le FCAR lui a accordé le

[1]Communication présentée lors de la Deuxième Conférence des chefs d'État et de Gouvernement de pays ayant en commun l'usage du français, sous le thème : Industries de la langue, Québec, 5 septembre 1987.

statut d'Organisme de service à la recherche (OSR). L'aide financière accordée dans le cadre de ce programme vise à permettre au Centre de fournir des services en ATO à tout chercheur subventionné, quelle que soit son affiliation au Québec. Cela détermine la seconde activité du Centre : d'une part, l'accueil, la formation et l'échange entre professeurs, chercheurs et étudiants grâce à des activités de communication (séminaires, colloques), des stages, la formation et le séjour au Québec de chercheurs étrangers; d'autre part, la production de services techniques, le traitement de données, le soutien à l'élaboration de dictionnaires, de modèles d'exploration de données textuelles et d'analyse-expertise sur les projets en ATO.

Enfin, depuis avril 1986, le Centre s'est vu reconnaître le statut de laboratoire par l'université. Ce statut a été obtenu sur la base de l'existence d'une masse critique de professeurs-chercheurs oeuvrant dans un domaine de recherche unifié par une certaine problématique commune. Au Centre, il s'agit de l'intérêt porté aux stratégies de description et d'exploration textuelles appliquées à des textes divers. Ainsi, l'originalité du Centre tient à son caractère véritablement inter-disciplinaire. Il existe, en effet, une très grande interdépendance entre les aspects informatiques, linguistiques, cognitifs et discursifs dans le traitement et l'analyse du matériel textuel. Le Centre tente de soutenir cette complexité par des environnements informatiques «paramétrables» par l'usager, permettant ainsi d'accueillir plusieurs théories ou hypothèses portant sur l'analyse textuelle.

1.3 *L'objet*

Le concept d'«analyse de textes par ordinateur» semble parler de lui-même. Pourtant, cette question en soulève bien d'autres. Nous avons choisi cette locution parce qu'elle désignait l'objet de notre recherche, le texte, et l'outil qui serait mis en oeuvre pour en produire l'analyse, l'ordinateur. Mais le domaine englobant de notre activité est celui de la compréhension des langues naturelles. Nous avons choisi de circonscrire notre recherche à la langue française, et plus particulièrement à l'étude des faits de langue dans les textes. Nous avons retenu comme objet le texte, qui n'est qu'un des supports de l'expression langagière, parce qu'il se prête plus facilement à des manipulations informatiques en raison même de son caractère plus formel. Il s'agit tout naturellement de la première étape dans la compréhension automatique du langage. D'un autre côté, le texte représente également une difficulté certaine dans la mesure où on procède à son décryptage extensif dans des contextes réels. En somme, notre objectif est de développer des systèmes de compréhension de la langue applicables à des objets réels.

Comprendre automatiquement la langue signifie qu'il faut «programmer des machines capables d'en analyser les diverses composantes : morphologique, syntaxique, sémantique, discursive et même pragmatique». Il faut donc «apprendre» à la machine une certaine compétence linguistique, mais aussi des compétences quant à l'organisation du discours, aux modalités de la représentation cognitive, de l'insertion sociale du langage, etc.

1.4 *Les domaines d'application*

En dehors des intérêts particuliers des chercheurs, on peut légitimement s'interroger sur la pertinence de l'ATO sur le plan social. Pour les besoins de la cause, on peut facilement distinguer deux grands domaines d'application de ces recherches : les aides à la lecture et les aides à l'écriture. Il existe de plus en plus de fonds documentaires dans les milieux de travail. On y manipule une très grande quantité d'information écrite à laquelle il n'est pas toujours facile d'accéder. Il s'avère intéressant de procurer une aide à la lecture de ce matériel. Voici quelques exemples. On peut souhaiter classifier l'information contenue dans des textes à partir des différents critères syntaxiques ou sémantiques qui permettraient, dans un deuxième temps, d'accéder à l'information par des procédures de repérage qui tiennent compte de ces critères. Ainsi on peut avoir accès à l'information non seule-

ment en partant de mots-clés, mais en recherchant des catégories préalablement repérées dans le texte. De telles procédures, simplifiées par des routines traduites sous forme de menus adressés à l'usager, sont faciles à définir dans l'état actuel de la recherche. On peut également appliquer des méthodes d'analyse de contenu des documents à partir des protocoles de description et d'exploration déjà développés dans des recherches plus fondamentales. L'utilisation de méthodes d'analyse de textes pour la construction assistée de banques de données relationnelles constitue un autre exemple d'application.

Les aides à l'écriture peuvent se révéler tout aussi précieuses. Pour la rédaction, on peut fournir un support lexical de nature orthographique en donnant accès à des dictionnaires, à un conjugueur ou encore, à un support lexical de nature sémantique en recourant à des banques de données analogiques ou terminologiques. Le document rédigé, on a la possibilité d'intervenir de nouveau en fournissant un correcteur orthographique et grammatical, ou en permettant un examen stylistique de l'écrit, par exemple en fournissant des indications sur le niveau de lisibilité. Un dernier type de système a trait au problème de l'annotation, c'est-à-dire à la préparation de résumés et de thésaurus ou à l'indexation automatique ou assistée des écrits.

2. Les environnements de programmation

2.1 *Trois systèmes*

Le Centre a regroupé les deux analystes qui sont à l'origine de programmes informatiques pour l'analyse de textes. Depuis sa reconnaissance officielle, le Centre a favorisé le développement de ces environnements qui constituent l'infrastructure informatique des différentes recherches qui y sont menées.

Le premier système, appelé «Système d'Analyse de Texte par Ordinateur» (**SATO**), a été développé principalement par François Daoust, d'abord au sein d'une équipe animée par Jean-Guy Meunier, ensuite en fonction des besoins spécifiques des chercheurs; il est maintenant développé au Centre d'ATO. Il s'agit d'un système interactif d'analyse de textes, spécialisé dans le repérage de concordances, la génération de lexiques et l'analyse statistique descriptive des données textuelles. Le système permet la manipulation non seulement des mots du texte, mais aussi des valeurs de propriétés, symboliques ou numériques, associées à ces derniers par l'usager. A partir d'une «syntaxe de requête», il est alors possible de produire des lexiques, des concordances et des dénombrements de tout ordre.

Le second système a été élaboré dans le cadre d'études de maîtrise et de doctorat, et de nouvelles versions développées en cours de recherche par Pierre Plante; ce dernier a été directement impliqué avec l'auteur dans la création du Centre. Le **Déredec** est un langage de programmation pour l'analyse du langage naturel. Il permet de représenter les données aux divers niveaux du fonctionnement linguistique : lexical, morphologique, syntaxique, sémantique et textuel. Il permet également de décrire ces données sous forme de grammaires s'adressant à ces divers niveaux. Il donne les moyens d'explorer ces diverses descriptions. Enfin, il est aussi utilisé pour la mise au point de systèmes à base de connaissances. Le **Déredec** est un cadre informatique général pour la programmation d'analyseurs ou simulateurs linguistiques basés sur diverses théories algorithmiques. On peut le considérer en premier lieu comme un atelier de travail pour les «informaticiens linguistes» et pour l'enseignement de leur discipline. Par ailleurs, il offre aussi toutes sortes de facilités pour l'utilisation des analyseurs linguistiques dans des applications telles que les analyses de contenu des textes et les systèmes-experts, qui questionnent les textes comme des banques de connaissances.

Grâce à ses ressources, le Centre permet également d'effectuer de la recherche en vue de la mise au point de systèmes-experts dans les domaines du diagnostic, de la classification et de la planification. Ce système, développé par Louis-Claude Paquin, est programmé en **Déredec** de telle sorte

que les structures utilisées pour la représentation informatique des objets cognitifs sont compatibles avec celles de tous les systèmes écrits en **Déredec**. **Déredec-EXPERT** constitue ainsi une structure de contrôle idéale pour les systèmes d'analyse de textes par ordinateurs.

2.2 *Les caractéristiques des systèmes*

Il est important de noter que les environnements de programmation développés au Centre sont nécessairement compatibles entre eux. **Déredec** offre une structure unique de représentation des données, les **EXFAD** (EXpressions de forme admissible au **Déredec**). Tout système écrit en **Déredec** fonctionne avec des **EXFAD** et produit à son tour des **EXFAD**. Cela permet une parfaite intégration de toutes les descriptions ou explorations produites en **Déredec** et rend possible leur utilisation d'une application à l'autre. Nous favorisons ainsi la double stratégie de modularisation et d'intégration. Nous développons les éléments des systèmes et les systèmes eux-mêmes sur une base modulaire, nous assurant ainsi de leur productivité et nous sommes dès lors en mesure de les intégrer sans modifier leur code. Il existe une interface entre **SATO** et **Déredec**, qui rend possible le transfert des données d'un système à l'autre.

2.3 *L'implantation*

Le **Déredec** est écrit en **Le_Lisp** et est compatible principalement avec Macintosh (Plus, SE, Mac II). Une version est aussi implantée sur VAX-VMS. **Déredec-EXPERT** est également compatible avec Macintosh (Plus, SE, Mac II). Enfin, **SATO** est compatible avec PC_XT-AT, VAX-VMS, Amdhal et est en développement sur Macintosh.

3. Les axes de développement et de recherche

Il faut distinguer les axes de recherche, fondamentale ou appliquée à des domaines, du développement proprement dit des environnements de programmation. On peut facilement résumer les orientations du Centre en trois volets : les deux premiers sont de nature fondamentale et le troisième, de la recherche appliquée. Nous appellerons les éléments du premier volet «Modules de Description Linguistique» (MDL) pour désigner l'ensemble des programmes destinés à la description proprement linguistique du texte avec l'objectif d'accroître le caractère automatique de la compréhension de la langue. Nous regroupons dans le second volet les «Modules de Représentation des Connaissances et d'Inférence» (MRCI). Il s'agit des systèmes consacrés à la construction de bases de connaissances relationnelles sous forme de structures arborescentes et pouvant être l'objet de l'application de procédures descriptives ou exploratoires en **Déredec**, ainsi que des modules d'inférence, comme on en rencontre dans les systèmes-experts. Le troisième volet, qui regroupe les «Modules d'Analyse de Textes» (MAT), renvoie aux deux premiers parce qu'il utilise leurs modules en les combinant dans un système intégré d'analyse de contenu assisté par ordinateur.

3.1 *Modules de description linguistique*

CBSF

Le premier module est un algorithme de «Catégorisation de Base Syntaxique du Français (**CBSF**), développé au Centre par Lucie Dumas. Il s'agit d'un progiciel qui reconnaît automatiquement la catégorie syntaxique de la plupart des formes lexicales de la langue française. Il constitue une

étape préalable à l'utilisation du progiciel de lemmatisation et de caractérisation morphologique dont il sera question ci-après. Il est également nécessaire au développement ou à l'application d'analyseurs syntaxiques. **CBSF** travaille à partir d'un ensemble de règles morphologiques et d'un dictionnaire d'exceptions. Une analyse morphologique permet d'associer la terminaison graphique d'unités lexicales à leur catégorie syntaxique respective. Par exemple, le suffixe **-able** correspond à des qualificatifs dans 1 380 des 1 432 formes qui se terminent ainsi en français. Notre dictionnaire répertorie 52 exceptions à cette règle, pour lesquelles il s'agit soit d'un nom, soit d'un nom/adjectif, soit encore d'un nom/verbe conjugué. Il existe 250 règles semblables et le dictionnaire d'exceptions comprend 3 000 entrées. Ce système arrive à catégoriser entre 85% et 90% des mots d'un texte, selon son niveau de difficulté. Les mots restants reçoivent, si désiré, leur catégorie en interactivité avec un opérateur humain. Les travaux actuels visent à réduire cette proportion de mots qui échappent à la catégorisation et il est permis d'espérer une performance de l'ordre de 95%.

LCMF

Le progiciel de «Lemmatisation et Caractérisation Morphologique du Français» (LCMF) a été également développé par Lucie Dumas. Il s'agit d'un programme permettant de regrouper automatiquement autour d'un même lemme toutes les formes flexionnelles qui y sont associées. La notion de lemme dans **LCMF** se définit comme l'unité minimale de représentation pour une forme lexicale donnée, telle qu'elle se présente normalement dans un dictionnaire. Ainsi, par exemple, toutes les désinences verbales sont représentées par l'infinitif et les formes fléchies des noms et des adjectifs, par le masculin singulier (ou le féminin singulier pour les défectifs). Les adverbes sont représentés par l'adjectif masculin singulier leur correspondant (ex. la forme lemmatisée de *brillamment* sera *brillant*). En plus de ce travail sur la lemmatisation, **LCMF** fournit également une catégorisation morphologique (genre, nombre, personne et temps) pour environ 90% des unités lexicales d'un texte. Ce pourcentage sera augmenté dans une version ultérieure.

La figure 1 illustre les éléments d'information obtenus à l'aide de **CBSF** et de **LCMF** sur la phrase : *Pour les réacteurs à neutrons, on dressera un bilan négatif*. On y retrouve les catégories syntaxiques pour chaque mot, le lemme pour les noms, le qualificatif et le verbe, ainsi que la catégorisation morphologique pour chaque forme lemmatisée.

FIGURE 1

Sortie de CBSF et de LCMF

c22 Pour	t1l les	MP réacteur nl2 réacteurs	c22 à	MP neutron nl2 neutrons	n211 on	IF3S dresser vl0 dressera	d12 un	MS bilan nl2 bilan	MS négatif dl3 négatif

c22:préposition forte
t1l: article ou clitique
nl2: nom commun
n211: pronom personnel
vl0: verbe conjugué
d12: déterminant nominal
dl3: adjectif qualificatif

MP: masculin pluriel
IF3S: indicatif futur, 3ème pers.
MS: masculin singulier

ALSF

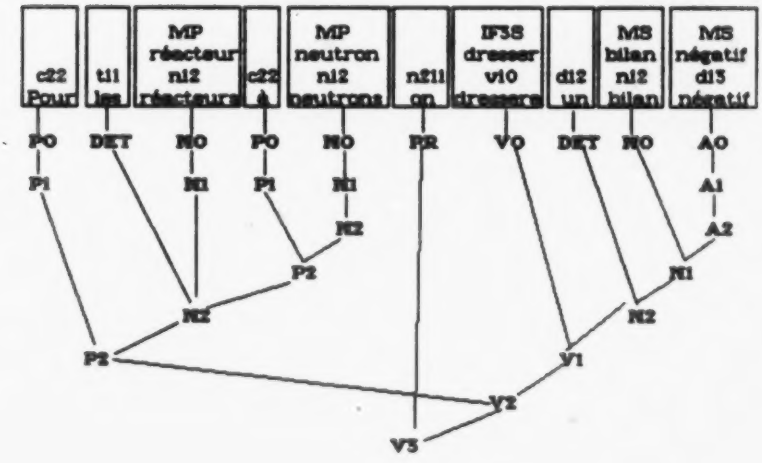
«L'Analyseur Lexico-Syntaxique du Français (**ALSF**) est un programme d'analyse automatique de français, développé dans le cadre d'un projet de coopération franco-québécois de recherche universitaire. Jean-Marie Marandin, de l'Institut national de la langue française (**INaLF**), travaille sur la

partie linguistique et Pierre Plante, du Centre d'ATO, soutient la programmation. Cet analyseur construit les structures syntagmatiques projetées par les catégories majeures du français : les noms, les verbes, les adjectifs et les prépositions. Il construit également les relations qu'entretiennent entre elles ces catégories dans des unités séquentielles : les propositions, les phrases et les segments du texte. La stratégie de développement consiste à favoriser une complète modularité, théorique et computationnelle, favorisant l'application effective des niveaux de description atteints et permettant l'expansion de la coquille initiale. Il s'agit d'un analyseur qui doit exécuter son traitement en regard de textes effectifs sans bloquer et sans produire d'analyses erronées pour les structures qu'il ne peut traiter. Des modules additionnels devraient pouvoir être développés en vertu de différentes hypothèses linguistiques. Écrit en Déredec, ALSF s'appuie sur les sorties CBSF et LCMF et comportera des modules de surcatégorisation.

La figure 2 présente la description syntaxique produite par ALSF de la même phrase que celle de la figure 1.

FIGURE 2

Sortie d'ALSF



3.2 *Modules de représentation des connaissances et d'inférence*

SAGA

Le «Système d'Analyse Grammaticale Assistée» (SAGA) est avant tout un système permettant la construction d'une structure de représentation des connaissances et l'instanciation de cette structure à partir d'objets réels. Un SAGA se présente sous la forme d'un arbre de décision contenant des branches et des feuilles et servant à représenter des connaissances sur un objet donné. Le SAGA comporte deux étapes. La première est celle de l'écriture. L'utilisateur construit sa structure de représentation à partir d'un système assisté qui lui propose des choix dans des menus et qui construit pour lui les automates et les modèles constitutifs de ce SAGA. Construire un SAGA ne consiste qu'à énoncer dans l'ordre hiérarchique les noms des membres et la nature (conjonctive ou disjonctive) de leur domaine, les liens pouvant relier les membres, ainsi que le type de feuille pouvant pousser à la fin des branches, c'est-à-dire à fournir le type d'information qui servira à l'instanciation. L'utilisateur d'un SAGA n'exécute que deux types d'opérations : il aiguillonne l'algorithme à chaque disjonction de membres et il remplit les feuilles des branches.

Le SAGA construit, la deuxième étape consiste à l'appliquer à des cas concrets. Le système assiste alors l'utilisateur dans l'instanciation à l'aide de menus. Il y a donc deux types d'arbres produits par SAGA : l'arbre général de décision, construit lors de la conception du système et l'arbre instancié, reflétant les choix effectués lors de l'application du SAGA à des objets réels.

Les SAGA s'appuient sur une stratégie inverse de celle des analyseurs dont le principe même est l'automaticité des procédures. Ici, les seuls automatismes sont contenus dans l'aide à l'écriture et à l'instanciation des SAGA. Autrement, toutes les décisions sont prises par l'usager au moment de l'exécution d'un SAGA.

Les SAGA permettent de décrire des objets complexes à partir de nombreuses dimensions. Ces objets peuvent aussi bien être des comportements observés dans une garderie, des comportement délinquants consignés dans des dossiers de la cour ou encore, des contenus d'entrevues semi-dirigées. Ce système permet de décrire ces objets et de définir les relations entre ses diverses dimensions. Toute la panoplie des automates ou des modèles d'exploration **Déredec** peut être appliquée à ces structures (**EXFAD**) écrites en **Déredec**. De plus, des stratégies d'inférence peuvent être associées à l'analyse des résultats obtenus des modèles d'exploration appliquée à ces bases de connaissances.

Les figures 3 et 4 présentent respectivement l'exemple d'un arbre de décision et d'un arbre instancié. L'exemple est tiré d'un SAGA écrit pour enregistrer l'information à propos d'une inscription quelconque dans un dossier judiciaire[2]. En général, chaque domaine structurant d'un SAGA peut être soit une conjonction (+) de membres, soit une disjonction (/) de membres. Dans l'exemple donné, une inscription est composée d'une conjonction de membres obligatoires (c'est-à-dire dont la présence est obligée) ou facultatifs (c'est-à-dire dont la présence n'est pas obligatoire). Dans la figure 3, les membres facultatifs se voient accolés d'un point d'interrogation (?). Certains membres peuvent être reliés à d'autres SAGA. Ainsi, le membre «acteur principal» est relié aux SAGA «fiche-acteur» et «fiche-tranche de vie». Cela implique la capacité du système d'établir des relations entre des bases de connaissances différentes. Les membres peuvent être soit une branche (comportant elle-même d'autres membres, comme le membre «intervention»), soit une feuille, comme «acteur principal», «juge», etc. (Chaque branche comporte une ou plusieurs feuilles à son extrémité).

Une feuille sera une expression atomique (valeur nominale), une quête au clavier, une quête au clavier avec apprentissage (confection d'une liste de valeurs), une quête dans une base de données, une quête dans un document (texte) ou encore un réseau de feuilles.

La figure 4 nous montre un exemple d'instanciation du SAGA «Inscription» à propos d'un cas. Il s'agit d'une dénonciation par l'infirmière de l'école, Garde Nantel, d'une jeune fille, Mademoiselle Turcotte, le 21 avril 1964. Le juge Lalonde sera assigné au dossier et la conséquence de cette dénonciation sera une comparution à venir à la cour de Montréal.

Déredec-EXPERT

Nous avons déjà défini le **Déredec-EXPERT**. Il faut retenir ici que ce générateur de systèmes-experts peut être utilisé à partir des représentations cognitives produites par SAGA. Cela permet d'appliquer la logique de l'inférence dans le traitement de l'information qui est contenue dans les SAGA. Il est d'ores et déjà prévu que le moteur d'inférence de **Déredec-EXPERT** sera utilisé dans le cadre de systèmes d'analyse textuelle.

---

[2]Cet exemple est tiré de la recherche : «Délinquance juvénile et psychiâtrie, 1910-1945», dirigée par Danielle Laberge et Jules Duchastel, financée par le CRSH.

FIGURE 3

Arbre de décision : fiche-inscription

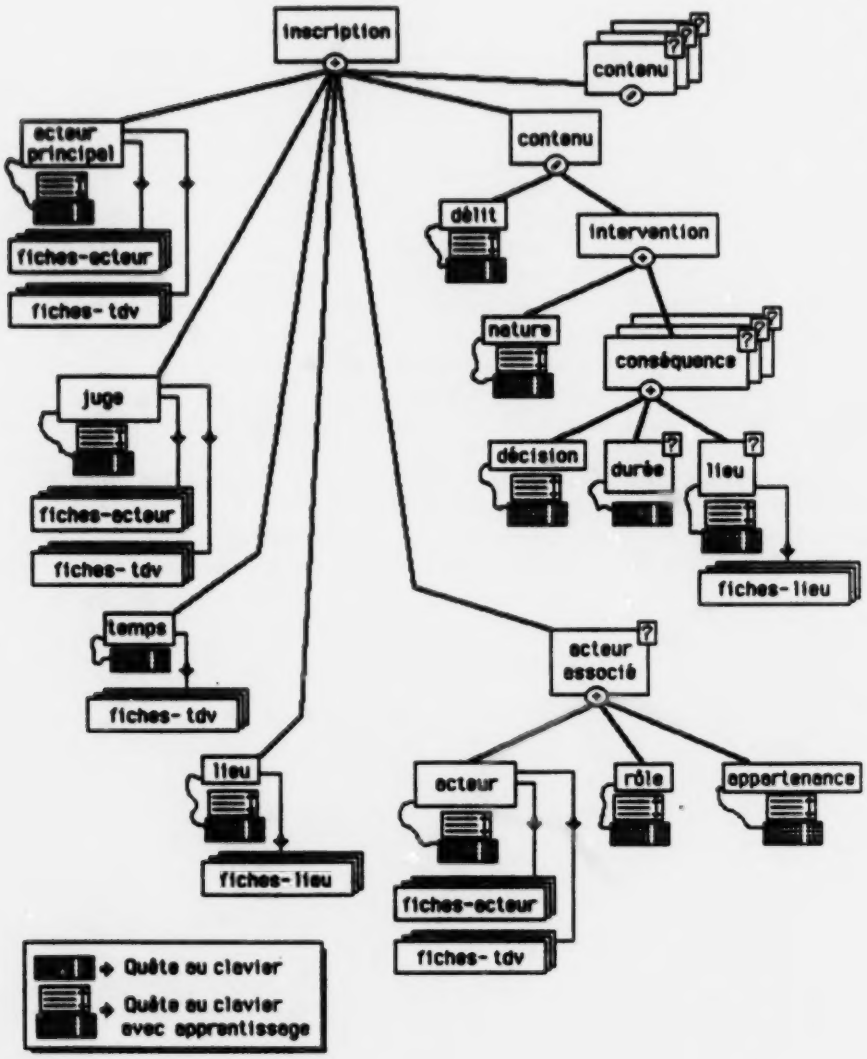
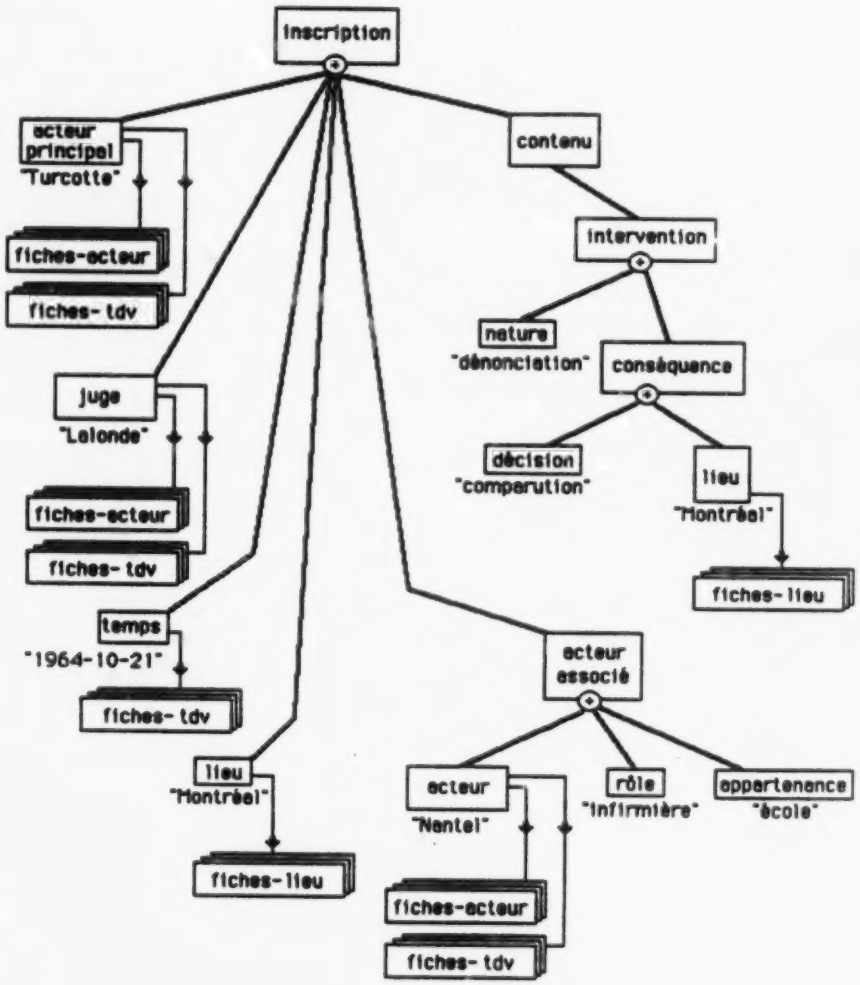


FIGURE 4

Arbre instancié : fiche-inscription



3.3 *Modules d'analyse de textes*

SACAO

Le «Système d'Analyse de Contenu Assistée par Ordinateur» (SACAO) vise à répondre au problème de l'aide informatisée à la lecture de données textuelles. Devant le volume considérable de textes qu'un nombre croissant de personnes sont appelées à décrypter dans le cadre de leur activité professionnelle et en raison du caractère peu systématique des méthodes existantes, nous voulons généraliser les applications découlant de certains modules de description, d'exploration et d'analyse textuelles développés au Centre.

Les procédures de lecture de SACAO sont destinées à des applications particulières. Tout en conservant leur valeur d'application générale, elles permettent ainsi un transfert méthodologique entre différents contextes. En somme, il s'agit d'un système intégré de lecture de textes comportant une

variété de modalités descriptives et exploratoires. La compatibilité de l'ensemble des modules écrits en **Déredec** et l'interface avec **SATO** permettent l'intégration des modules de description (morphologique, syntaxique, sémantique, cognitive, textuelle), d'exploration des données décrites et d'analyse des résultats. Le but recherché est de permettre à l'utilisateur du système de paramétrer les procédures en fonction de son approche, soit en lui laissant le choix des modules pertinents pour son objet, soit en lui permettant de définir certains paramètres particuliers de nature lexicale, morphologique, sémantique ou cognitive pour l'un ou l'autre modèle, soit en lui fournissant des options différentes pour l'analyse des données.

Outre la mise au point des procédures d'intégration et de modularisation du système, une attention particulière est portée à trois axes de développement de systèmes d'analyse de contenu. Nous voulons ainsi incorporer des modules de descriptions textuelles allant au-delà de la syntaxe ou de la morphologie. Il existe présentement des prototypes d'analyse de l'argumentation d'un texte ou encore, de segmentation discursive. Nous construisons également de nouvelles procédures de dictionnaires sémantiques structurés mettant en oeuvre les possibilités de programmation offertes par **SAGA**. Nous élaborons enfin des procédures pour le repérage et l'analyse de la référence dans le discours (étude de la citation, du discours rapporté, du discours indirect libre).

CORRECTEUR ORTHOGRAPHIQUE DÉVELOPPÉ AU QUÉBEC[1]

Richard Kromp
IBM Canada

Dans cette communication, nous décrirons l'expérience d'**IBM Canada** dans le développement d'un correcteur orthographique du français. Il s'agit de **SCOT**, acronyme pour Système de correction orthographique et textuelle. Après la description de l'historique du projet, nous donnerons un aperçu de la contribution des services linguistiques.

Étude du projet

Le point de départ du projet est un système américain d'aide à la rédaction appelé **PROOF**. Ce système identifie les fautes d'épellation dans un texte anglais de nature générale, propose des mots bien orthographiés en tenant compte des variantes orthographiques entre l'usage américain et l'usage anglais, repère certaines expressions maladroites ou superflues et isole les mots qui dépassent un niveau de langue donné, en fonction des niveaux de scolarité. Par exemple, il fait la distinction entre *teacher*, d'un usage commun, et *tutor*, beaucoup plus relevé. Le système était destiné à être intégré aux progiciels de traitement de textes pour venir en aide aux employés appelés, dans le cadre de leur fonction, à faire de la correspondance commerciale.

Des pourparlers entre le Centre de recherche de Gaithersburg qui avait développé **PROOF** et les services linguistiques d'**IBM Canada** ont commencé en avril 1982. Le responsable du projet (le Dr Rosenbaum, directeur du groupe Technologie bureautique de pointe) était alors à la recherche d'une équipe de linguistes capables de mettre au point le pendant français du produit.

Après les discussions d'usage, nos services linguistiques obtenaient le mandat de développer le système.

L'équipe

L'équipe de recherche était constituée de six spécialistes canadiens. Il s'agissait d'une équipe multidisciplinaire formée de quatre employés d'**IBM** (Johanne Boucher, traductrice, Vivianne Cohen, directrice de la traduction, Marylène Le Deuff, documentaliste et Gaston Robitaille, informaticien). Une étudiante en linguistique de l'université de Montréal, France Boucher, a alors été engagée sur une base contractuelle, et le Dr William Mackey, lexicographe de l'université Laval, servait de consultant à l'équipe.

---

[1] Communication présentée lors de la Deuxième Conférence des chefs d'État et de Gouvernement de pays ayant en commun l'usage du français, sous le thème : Industries de la langue, Québec, 5 septembre 1987.

Étapes

Le développement s'est fait en trois étapes :

1- recherche documentaire (pour faire le point sur l'état de la question);
2- établissement du vocabulaire;
3- tests.

Phase I - recherche documentaire

L'objectif de la recherche documentaire était d'identifier les publications existantes en français, sur lesquelles le travail de l'équipe pouvait reposer. A la lumière de ces résultats, il fallait ensuite identifier les fonctions du programme américain qui pouvaient (ou devaient) être mises au point en français.

La documentaliste a établi une bibliographie détaillée des articles, études, rapports, thèses, dictionnaires (techniques et généraux), grammaires, guides de correspondance disponibles jusqu'alors. De l'examen de toute cette documentation l'équipe a pu tirer quatre conclusions.

1- Personne n'avait jusqu'alors produit d'étude de la langue française par rapport aux niveaux de scolarité, comparable à ce qui existe pour l'anglais.
2- On n'a pas trouvé d'études de fréquence du vocabulaire français portant essentiellement sur la correspondance commerciale moderne (d'où la nécessité de constituer un corpus original).
3- Contrairement à l'anglais britannique et l'anglo-américain, il y a peu de différences entre l'usage orthographique en français hexagonal et en français québécois. Les différences lexicographiques se situent principalement sur le plan sémantique. Si aux yeux du québécois, le mot *cabaret* évoque un plateau, il signifie pour le français *nightclub* ou *dancing*. Mais d'une rive à l'autre de l'Atlantique, on s'entend sur sa graphie. Les exceptions sont relativement banales; il y a bien quelques accents sur des majuscules qui flottent un peu plus dans les interlignes hexagonaux, mais les différences s'arrêtent là.
4- Finalement, la documentation permettant de réaliser pour le français les fonctions du programme américain existe bel et bien, mise à part l'identification du vocabulaire par niveau de scolarité. Qui plus est, le nombre de dictionnaires et de lexiques spécialisés est impressionnant. Le projet visera donc d'abord à développer les fonctions de correction orthographique, les fonctions textuelles devant faire l'objet de recherches subséquentes.

Phase 2 Constitution du vocabulaire

Il fallait, au point de départ, constituer un corpus original, compte tenu du contexte commercial dans lequel le correcteur allait être utilisé. Il fallait dépouiller la correspondance d'affaires. Pour ce faire, IBM bénéficiait de milliers de documents stockés sur médias électroniques (disquettes, cartes et bandes magnétiques). Ces documents avaient déjà été recueillis chez les clients d'IBM France en 1977. Les données apparaissant sous des formats multiples, il fallait d'abord les convertir en format standard. De ces textes, un corpus de 150 000 entrées françaises a été compilé et les entrées classées par ordre de fréquence. Par souci d'économie et de performance, les entrées devaient être inscrites sous forme de racines et de désinences. L'équipe a alors dû développer une structure de dérivation particulière au français pour tenir compte des spécificités de la langue française (déclinaisons verbales, formes adjectivales, etc.). Il a également fallu modifier les programmes américains pour qu'ils reconnaissent les caractères français accentués.

Une révision initiale s'imposait cependant, dans le but d'éliminer les mots mal orthographiés, certains hapax sans intérêt et des initiatives malheureuses de l'équipe américaine qui nous suggérait des formes complémentaires de type *hommes enceints* et *tu pleuvais*. Il fallait aussi ajouter à la liste lexicale certains mots du français fondamental malheureusement absents de l'échantillonnage (certains titres par exemple), des formes verbales et adjectivales complémentaires, des éléments tirés d'un corpus parallèle établi à Paris dans le cadre d'un projet de recherche de transcription sténographique.

Ajout des noms propres

L'équipe de recherche était alors relativement satisfaite de la liste établie, mais un problème subsistait toujours : la liste des noms propres était lacunaire.

Pour la compléter, on a ajouté les noms de lieux géographiques les plus usuels en France, au Canada et aux États-Unis, les noms des 500 plus grandes sociétés françaises et canadiennes, les prénoms les plus courants, les noms de famille occupant plus de deux pages dans l'annuaire téléphonique de Montréal et de Paris (comme Boucher, mais pas Kromp) ainsi que les noms de villes et de pays importants.

Une fois tout ce travail effectué, les fichiers ont été constitués, et la liste regroupe 20 000 mots-racines pouvant s'étendre à 205 000 mots individuels.

Phase 3 Test

Le produit a ensuite été testé, comme c'est le cas de tout produit en développement. On a «scotté» (si vous me permettez ce néologisme maintenant usuel à nos services linguistiques) 55 000 mots tirés de lettres-types trouvées dans huit guides de correspondance française, de la correspondance du service de communications de Montréal et des propositions rédigées par le personnel des succursales d'IBM à Montréal. Ces textes ont tous été saisis sans révision par une secrétaire. Le taux d'efficacité a été établi à 99,85%. Le produit pouvait donc être remis, dès août 1983, au Centre de recherche de Gaithersburg et intégré aux logiciels de bureautique IBM (par exemple, Visio 4 et les services personnels PC).

Ce résultat est le fruit d'un savoir-faire québécois. Il faut toutefois être honnête, et rendre à César ce qui lui revient : les programmes initiaux ont été conçus par des chercheurs américains. Mais le savoir-faire québécois conjugué à l'effort de la France en a permis l'adaptation.

DÉCOUPAGE AUTOMATIQUE DE TEXTES EN MOTS SIMPLES[1]

Jacques Ladouceur
Langues et linguistique
Université Laval

1. Introduction

En linguistique, l'analyse automatique implique souvent au moins deux éléments : un texte à analyser, dont la saisie a été faite à l'aide d'un logiciel de traitement de texte, et un logiciel d'analyse.

Sauf exception un texte, sous sa forme textuelle (suite de phrases), n'est pas compatible avec un logiciel d'analyse. Pour être en mesure d'être traité par un logiciel d'analyse, un texte doit subir une transformation qui consiste en un découpage en mots simples. Voici un texte avant (voir figure 1) et après un découpage (voir figure 2) :

FIGURE 1

Sans parler de la mer, des femmes ou du soleil, chaque caillou a son histoire. Chaque taillis sa forêt vierge, chaque ruine sa muraille de Chine, ses falaises d'Étretat, et le moindre coin de rue ses jardins suspendus.

FIGURE 2

sans	000001 000001
parler	000001 000002
de	000001 000003
la	000001 000004
mer ,	000001 000005
des	000001 000006
femmes	000001 000007
ou	000001 000008
du	000001 000009
soleil ,	000001 000010

[1]Communication présentée lors du 56e Congrès de l'ACFAS, Moncton, 9-13 mai 1988.

2. La problématique du découpage automatique

2.1 Caractéristiques particulières des textes.

Lorsqu'on observe attentivement un grand nombre de textes, du point de vue de la mise en page, on se rend compte qu'ils ont chacun des caractéristiques particulières. Rarement deux textes sont complètement identiques. Voici quelques exemples :

Ex. 1

{P.1}

[PARAGRAPHE 1]
SANS PARLER DE LA MER, DES FEMMES OU DU SOLEIL, CHAQUE CAILLOU A SON HISTOIRE. CHAQUE TAILLIS SA FORÊT VIERGE, CHAQUE RUINE SA MURAILLE DE CHINE, SES FALAISES D'ÉTRETAT, ET LE MOINDRE COIN DE RUE SES JARDINS SUSPENDUS.

Ex. 2

[P.1]

<LOUIS LAINE> DIS-MOI QUE TU M'AIMES ENCORE. LA NUIT EST VENUE! MAINTENANT JE SUIS LÂCHE! MAINTENANT JE PUIS PRONONCER DE TELLES PAROLES! <MARTHE> IL EST TROP TARD. TU N'ENTENDRAS POINT LE MOT QUE TU DEMANDES DE MA BOUCHE. SONGE À TOI SEUL! <LOUIS LAINE> EH BIEN DONC, MALHEUR À MOI!

Ex.3

[P.1]

LOUIS LAINE	DIS-MOI QUE TU M'AIMES ENCORE. LA NUIT
LOUIS LAINE	EST VENUE! MAINTENANT JE SUIS LÂCHE!
LOUIS LAINE	MAINTENANT JE PUIS PRONONCER DE TELLES
LOUIS LAINE	PAROLES!
MARTHE	IL EST TROP TARD. TU N'ENTENDRAS
MARTHE	POINT LE MOT QUE TU DEMANDES DE MA
MARTHE	BOUCHE. SONGE À TOI SEUL!
LOUIS LAINE	EH BIEN DONC, MALHEUR À MOI!

Dans le premier exemple, on constate la présence d'un numéro de page en haut à droite. Le numéro de page est encadré par les signes { et }. On constate aussi la présence d'un numéro de paragraphe qui est encadré par les signes [et].

Dans le deuxième exemple, un extrait de pièce de théâtre, il n'y a pas de numéro de paragraphe. Cependant, on peut voir un élément supplémentaire : le nom de la personne qui parle. Cette information est placée entre les signes < et >. De plus, le numéro de page se situe en haut au centre et est placé entre les signes [et].

Dans le dernier exemple, le texte est le même que dans l'exemple précédent. Toutefois, il y a des différences. Le nom de la personne qui parle n'est encadré par aucun signe. Il est placé à des endroits fixes de chaque ligne de texte à l'intérieur d'une zone de quatorze caractères.

2.2 Exigences particulières d'un logiciel d'analyse

D'une façon générale, tous les logiciels d'analyse ont chacun des exigences particulières. Ces exigences portent principalement sur deux plans. D'une part, un texte découpé doit être placé dans un fichier qui a une structure adaptée au logiciel d'analyse. D'autre part, le fichier doit contenir tous les éléments d'information que le logiciel d'analyse exige (ex. numéro de ligne, ordre du mot dans la ligne, numéro de paragraphe, etc.) et ces éléments doivent être placés dans un ordre précis. Voici quelques exemples de structures[2] de fichiers que des logiciels d'analyse pourraient exiger :

Ex. 1

pendant		000001	000001

Ex. 2

PENDANT	P. 1	000001	00001	

Ex. 3

DIS	LOUIS LAINE	P.1	00001	1	F	F	0	-

Ex. 4

PENDANT	P.1	01	1	F	000

Dans le premier exemple, on voit que la fiche est composée de quatre zones. La première zone contient le mot, la seconde, un signe de ponctuation, la troisième, le numéro de la ligne d'où provient le mot et, finalement, la dernière contient l'ordre du mot dans la ligne d'où il provient. Remarquons aussi que le mot est en caractères minuscules.

Dans le deuxième exemple, la fiche est constituée de cinq zones. La première zone, qui est plus courte que dans l'exemple précédent, contient le mot, qui cette fois apparaît en caractères majuscules. La seconde zone contient un numéro de page. Les troisième et quatrième zones contiennent respectivement un numéro de ligne et l'ordre du mot dans la ligne où il est.

Dans le troisième exemple, la fiche est constituée de neuf zones. La première contient le mot, qui cette fois n'a été converti ni en caractères minuscules ni en caractères majuscules. La seconde zone contient le nom de la personne qui parle (dans le cas d'une pièce de théâtre, par exemple). La troisième zone contient un numéro de page. Les zones quatre et cinq contiennent un numéro de ligne et l'ordre du mot dans la ligne dont il provient. Remarquons aussi que la longueur de ces deux

[2]Les exemples que nous présentons sont fictifs. En général, les structures qu'exigent les logiciels d'analyse sont beaucoup plus complexes. Ces structures ont été créées afin de simplifier la présentation.

dernières zones n'est pas la même que dans les exemples précédents. Les trois zones suivantes contiennent des éléments d'informations qui ne sont pas liés à un texte mais qui peuvent être exigés par un logiciel d'analyse. Finalement, la neuvième zone contient un signe de ponctuation.

Le dernier exemple montre les fiches qui sont constituées de quatre zones. On peut voir que la zone réservée pour le mot est beaucoup plus longue que dans les exemples précédents et qu'il n'y a pas de zone réservée pour un signe de ponctuation.

2.3 Contraintes

Étant donné qu'un texte possède des caractéristiques particulières et qu'un logiciel d'analyse a des exigences précises, nous sommes confrontés à la situation suivante : un programme de découpage peut lire un texte de type X, et seulement de type X, et le transformer en X', et seulement en X'.

Lors de la mise au point d'un programme de découpage de textes, la forme des textes à saisir et la forme des textes après découpage doivent être établies. Cela entraîne un certain nombre de contraintes.

D'abord, les textes non conformes au programme de découpage qu'on désire utiliser doivent être remaniés. Ensuite, si pour une raison quelconque on décide de modifier la forme de nouveaux textes à saisir, on doit alors modifier ou refaire le programme de découpage. Bref, s'il n'y a pas de compatibilité entre un texte et un programme de découpage, on doit, soit modifier le texte, soit modifier le programme de découpage.

2.4 Solution proposée

La solution qui pourrait être apportée au problème de découpage est la suivante : un logiciel de découpage automatique de textes en mots simples pourrait lire un texte dont la saisie a été faite selon n'importe quelle convention et le transformer sous n'importe quelle forme.

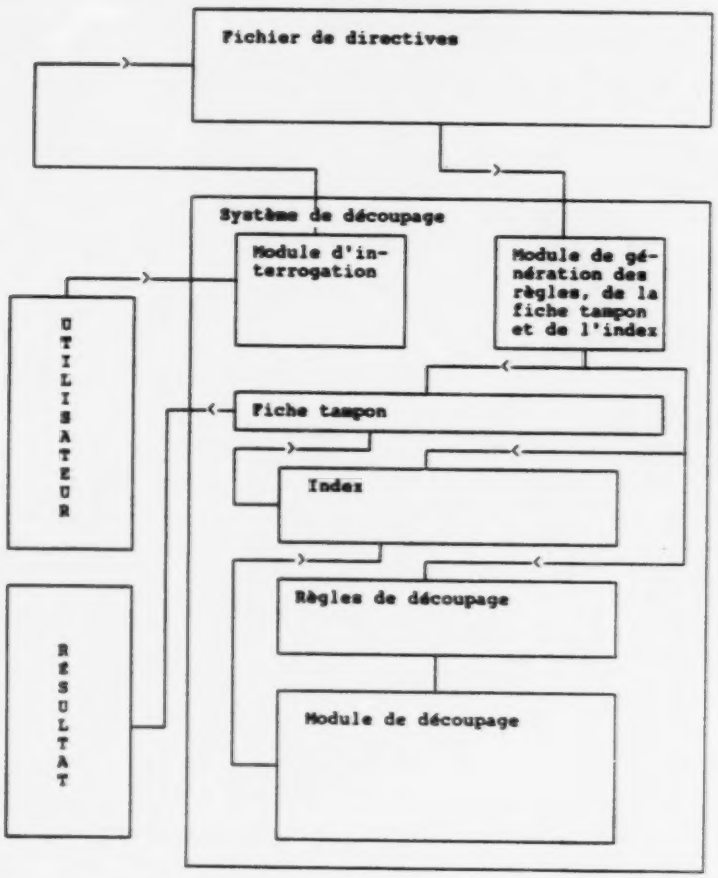
3. D.A.T., un logiciel de découpage automatique de textes

Lorsqu'on procède au découpage automatique d'un texte, on doit tenir compte d'une certaine quantité d'informations qui varient d'un texte à l'autre. De plus, on doit prévoir la possibilité de générer un résultat qui peut prendre une infinité de formes. Pour cela nous avons créé un système de découpage **D.A.T.**, qui fonctionne en utilisant les directives qui proviennent de l'utilisateur.

Les directives de découpage peuvent être obtenues à l'aide d'un module d'interrogation qui fait partie du système de découpage. Le module interroge l'usager afin d'obtenir une description du texte à découper et de la forme sous laquelle le texte sera transformé. L'information transmise par l'utilisateur est acheminée dans un fichier de directives. Au moment du découpage, le système de découpage active un module de génération. Ce dernier procède à une lecture du fichier de directives et génère une série de règles de découpage, une fiche tampon, qui correspond à la forme désirée pour le résultat, et finalement un index permettant d'acheminer correctement l'information tirée du texte à l'intérieur de la fiche tampon. A l'intérieur du système de découpage, il y a finalement un module de découpage qui fonctionne en utilisant les règles de découpage et qui transmet ses résultats via la fiche tampon et l'index à l'intérieur du fichier de sortie.

Voici le schéma du système de découpage :

Voici le schéma du système de découpage :

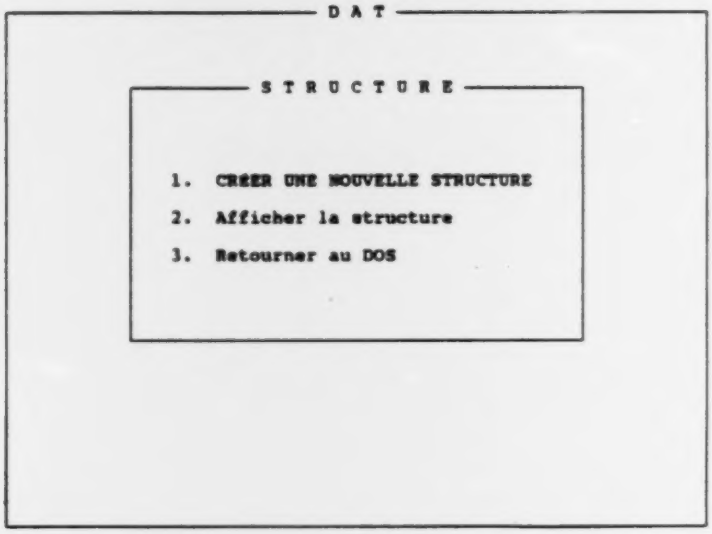


3.1 Le lancement du module d'interrogation

Pour activer le module d'interrogation, il suffit de fournir à l'ordinateur, à partir du D.O.S., l'instruction suivante : **DAT <nom du fichier structure>/s. DAT** est le nom du logiciel de découpage. **Le signe /s** indique au logiciel qu'on désire activer le module d'interrogation afin de créer ou modifier le fichier structure dont le nom est précisé juste avant le trait oblique. Par exemple, si on veut créer ou modifier le fichier structure *lemme.str* on donne l'instruction *DAT lemme.str/s*.

Lorsque le module d'interrogation a été activé, un premier menu apparaît à l'écran (voir figure 3).

FIGURE 3



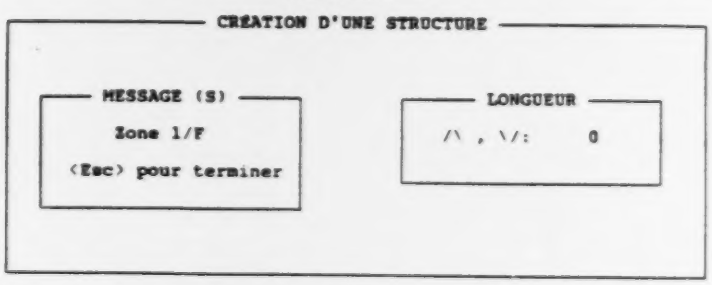
Le menu offre trois options. La première permet de créer un fichier structure lorsqu'il s'agit d'un fichier qui n'existe pas. Si le fichier est déjà créé, le module d'interrogation choisit automatiquement *1- MODIFIER UNE STRUCTURE*. On peut choisir l'une ou l'autre des trois options à l'aide des touches flèches haut et flèches bas du clavier. L'option sélectionnée apparaît en caractères majuscules. Par exemple, si on appuie sur la touche flèche bas l'option 1 se réécrit en caractères minuscules et l'option 2 se réécrit en caractères majuscules.

3.2 La création d'une structure

Si on choisit la première option, *créer une nouvelle structure*, alors l'écran affiche un tableau qui s'intitule *CRÉATION D'UNE STRUCTURE* (voir figure 4).

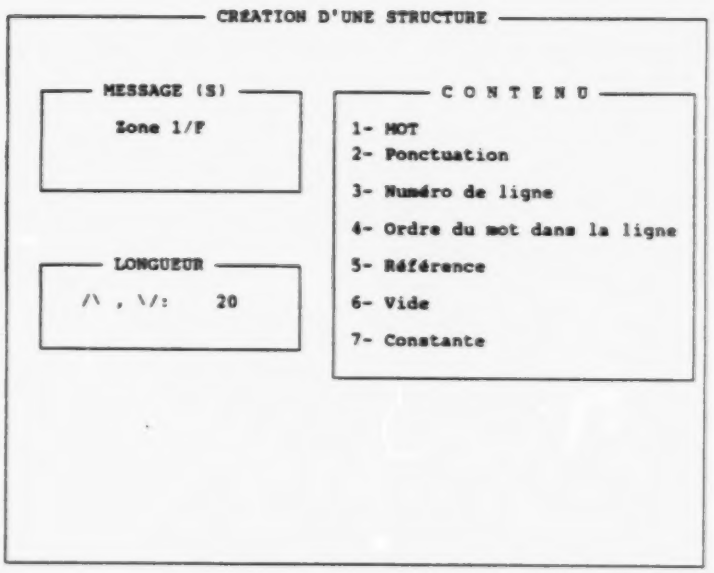
Il y a deux fenêtres à l'intérieur du tableau. La première s'intitule *Message (s)* et la seconde *Longueur*. Dans la fenêtre *Message (s)* deux indications sont inscrites. La première, *Zone 1/F*, montre à l'utilisateur qu'il décrit la première zone et que cette zone est la dernière de la fiche. Il est tout à fait normal que la première zone soit la dernière puisque l'étape de création commence. Si l'utilisateur était en train de modifier une structure composée de huit zones il verrait affiché le message *Zone 1/8*. Le second message indique à l'utilisateur qu'il peut mettre fin à l'étape de création d'une structure en appuyant sur la touche d'échappement Esc.

FIGURE 4



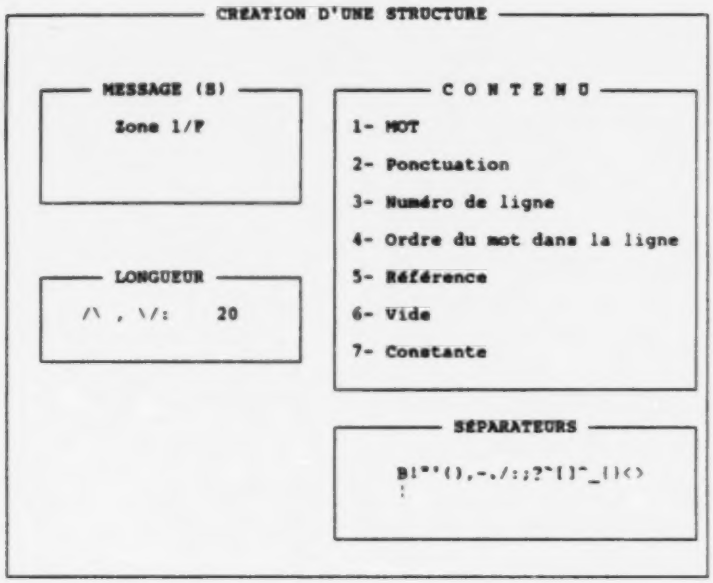
La seconde fenêtre, *longueur*, indique à l'utilisateur qu'il doit préciser la longueur de la zone indiquée dans la fenêtre des messages. Dans l'exemple, il s'agit de la zone 1. La présence des deux flèches sert à indiquer qu'on peut faire augmenter ou diminuer le chiffre, placé à droite des flèches, en utilisant les touches **flèche bas** et **flèche haut** du clavier. Le chiffre correspond à la longueur de la zone. Lorsqu'il détermine la longueur de la zone, l'utilisateur appuie sur la touche **retour** et le module d'interrogation modifie le tableau. La fenêtre **longueur** se déplace sous la fenêtre des messages. Ensuite, une nouvelle fenêtre qui s'intitule *contenu* apparaît (voir figure 5).

FIGURE 5



En faisant apparaître la fenêtre **contenu**, le module d'interrogation indique à l'utilisateur qu'il doit préciser le contenu de la zone dont le numéro apparaît dans la fenêtre des messages. Sept types de contenus sont possibles : un mot, un signe de ponctuation, un numéro de ligne, l'ordre du mot dans la ligne, une référence, un vide ou une constante. L'utilisateur peut choisir un des types en utilisant les touches **flèche bas** et **flèche haut** du clavier. Le choix apparaît toujours en caractères majuscules. Lorsque le contenu a été précisé, on appuie sur la touche **retour**. Nous choisirons, en guise de contenu, l'option **mot**. Une nouvelle fenêtre apparaît, *séparateurs*, sous la fenêtre **contenu** (voir figure 6).

FIGURE 6



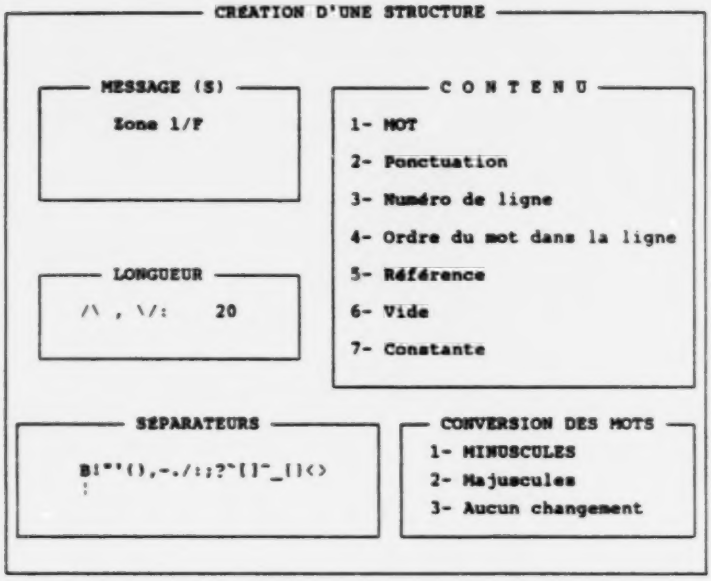
En faisant apparaître la fenêtre **séparateurs** le module passe au second plan d'interrogation et indique à l'usager qu'il a besoin d'un supplément d'information. En d'autres mots, il demande à l'utilisateur comment le module de découpage devra faire pour reconnaître les mots dans le texte qu'il découpera. Sachant à l'avance que le découpage des mots simples se fait en consultant une liste de séparateurs, il présente donc une liste générale de séparateurs que l'utilisateur peut modifier. L'utilisateur peut, par exemple, sélectionner un séparateur à l'aide des touches **flèche droite** et **flèche gauche** ou le faire disparaître avec la touche **flèche bas** ou encore, le faire réapparaître avec la touche **flèche haut**.

Lorsque l'utilisateur a sélectionné les séparateurs qu'il veut utiliser, il appuie sur la touche **retour**. La fenêtre **séparateurs** est déplacée vers la gauche pour faire place à une nouvelle fenêtre qui s'intitule **conversion des mots** (voir figure 7).

La fenêtre **conversion des mots** permet à l'utilisateur d'indiquer s'il désire que les mots soient convertis en caractères minuscules, majuscules ou qu'ils demeurent tels qu'ils apparaissent dans le texte. Lorsqu'il a fait son choix, l'utilisateur doit appuyer sur la touche **retour**. Le module revient alors au premier plan d'interrogation et passe à la zone suivante (voir figure 8).

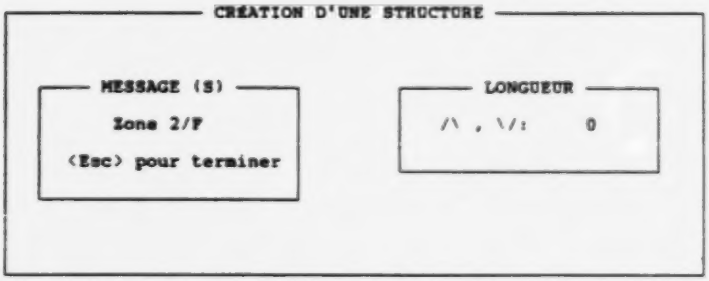
La fenêtre des messages indique qu'il faut décrire la zone 2. Si l'utilisateur désire une seule zone dans la fiche, il faut appuyer sur la touche **Esc** et le module d'interrogation met fin à la séance. Pour la démonstration, nous allons continuer et la longueur de la seconde zone sera d'un seul caractère (voir figure 9).

FIGURE 7



Maintenant, nous allons indiquer au module d'interrogation que le contenu de la zone est un signe de ponctuation. Pour cela nous choisissons la deuxième option dans la fenêtre contenu. Comme le module n'a pas besoin d'autre information, il ne passe pas au second plan d'interrogation. Il présente à nouveau la fenêtre longueur et indique, dans la fenêtre des messages, que la troisième zone devra être décrite (voir figure 10).

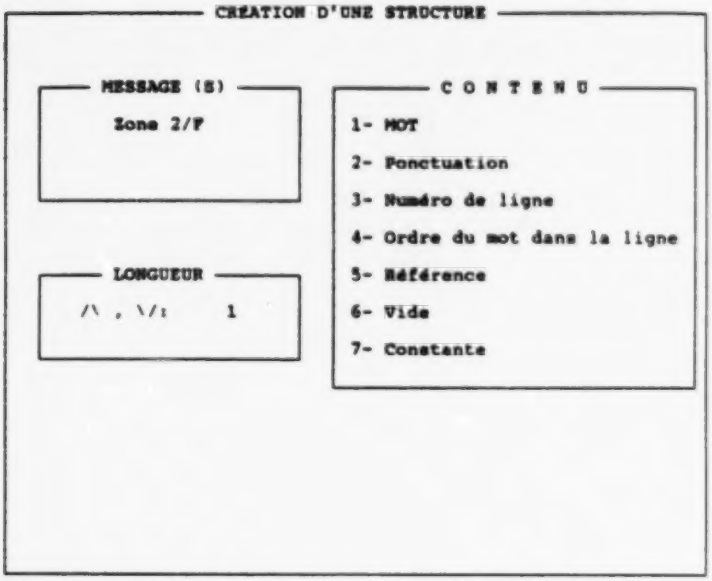
FIGURE 8



Lorsque le contenu d'une zone est un signe de ponctuation, un numéro de ligne, un numéro de mot ou un vide le module réagit toujours de la même façon. Il passe immédiatement à la zone suivante. Cela s'explique par le fait que le système de découpage n'a pas besoin d'indication afin de trouver ces éléments dans le texte à découper.

Lorsque l'utilisateur indique que le contenu d'une zone est une référence, une nouvelle fenêtre apparaît à l'écran (voir figure 11).

FIGURE 9



Lorsque le contenu d'une zone est une référence, le module s'informe à savoir s'il s'agit d'une référence à position variable ou d'une référence à position fixe. Pour répondre, il s'agit de sélectionner une option à l'aide des touches flèche haut et flèche bas. Le choix apparaît en caractères majuscules.

FIGURE 10

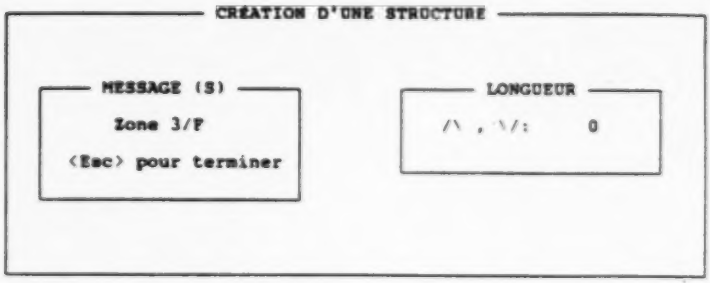
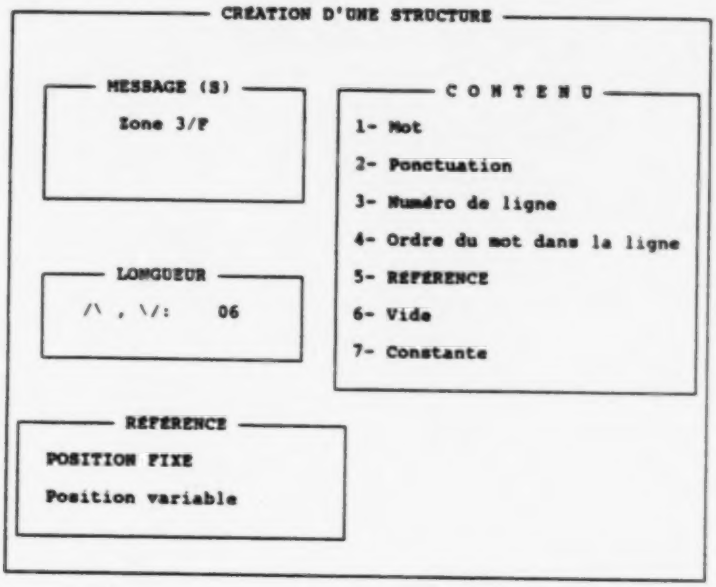


FIGURE 11

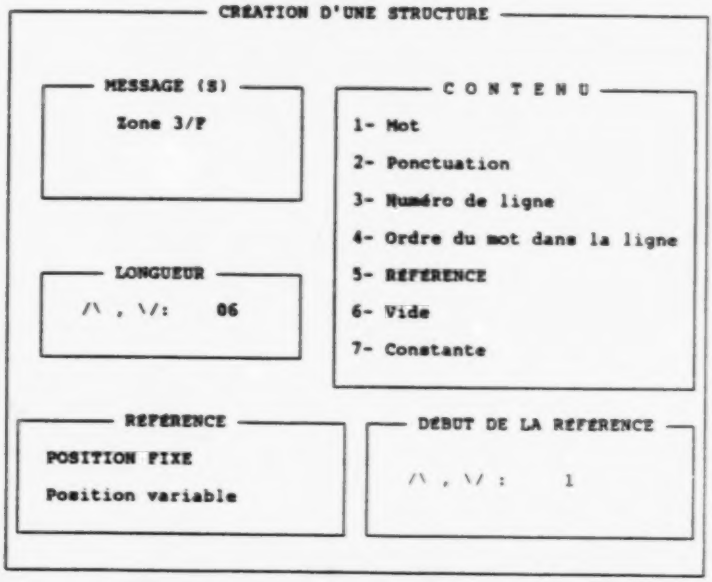


Si la référence est à position fixe, le module fait apparaître une nouvelle fenêtre qui s'intitule *début de la référence* (voir figure 12).

À cette étape, l'usager indique, en utilisant les touches **flèche haut** et **flèche bas** la position à laquelle commence cette référence à chaque ligne de texte. En guise d'exemple, choisissons la position 1. Lorsqu'on appuie sur la touche **retour** le module déplace la fenêtre vers la gauche, en la superposant à la fenêtre **référence**, et fait apparaître une nouvelle fenêtre qui s'intitule **longueur de la référence** (voir figure 13).

Ce faisant, le module indique à l'utilisateur qu'il désire connaître la longueur des références qui commencent à la position 1 à chaque ligne de texte. Nous choisirons une longueur de 6.

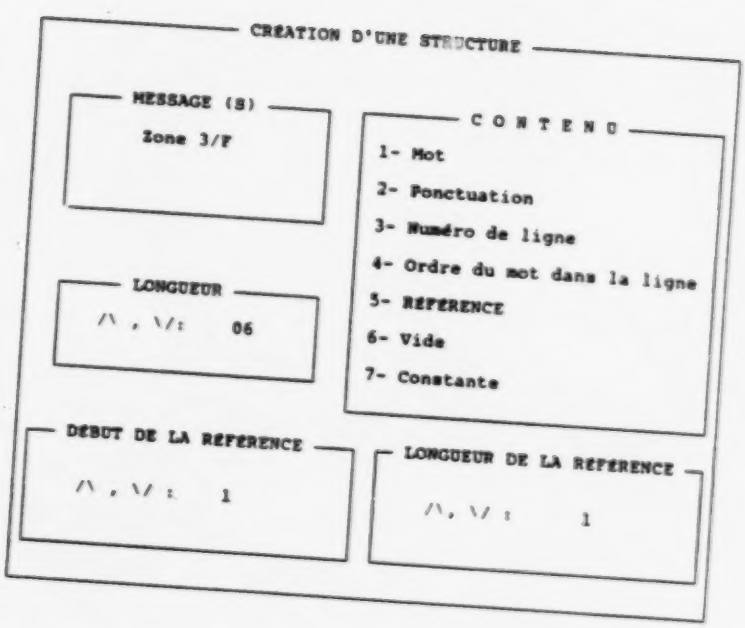
FIGURE 12



Lorsqu'on indique au module qu'une référence est du type **position variable** alors, à la place de la fenêtre **début de la référence**, il fera apparaître une fenêtre qui s'intitule **délimiteur gauche** (voir figure 14).

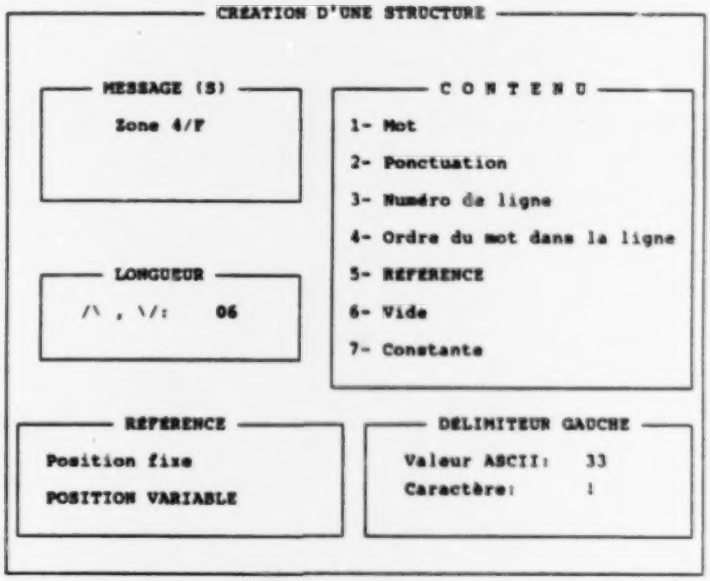
Deux éléments apparaissent dans la nouvelle fenêtre : une valeur ASCII et un caractère. La valeur ASCII est le code décimal ASCII qui correspond au caractère placé en dessous.

FIGURE 13



Lorsqu'on appuie sur la touche flèche haut, le code décimal augmente et le caractère correspondant au code apparaît en dessous. Lorsqu'on appuie sur la touche flèche bas, le code décimal diminue et le caractère correspondant au code apparaît en dessous. On peut donc choisir le caractère qui est utilisé dans le texte en guise de délimiteur gauche de la référence. Choisissons le caractère (qui a la valeur ASCII 40.

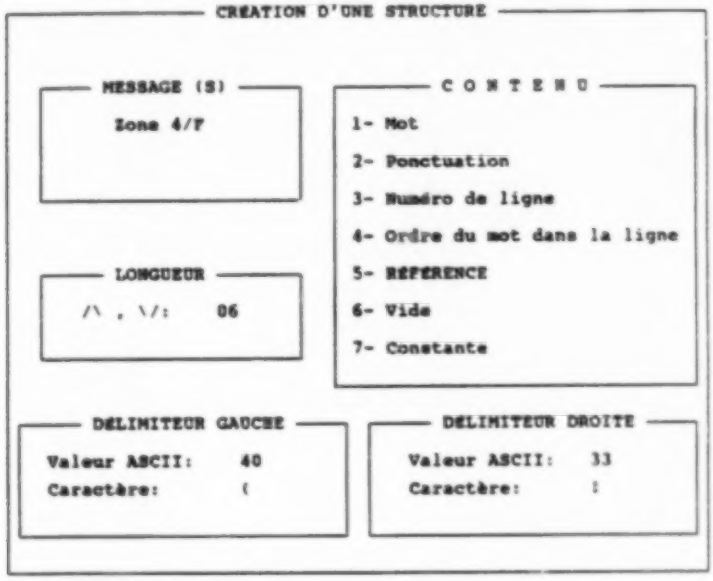
FIGURE 14



Lorsque le choix a été fait et qu'on appuie sur la touche retour, la fenêtre est déplacée vers la gauche pour faire place à une nouvelle fenêtre qui s'appelle *délimiteur droite* (voir figure 15).

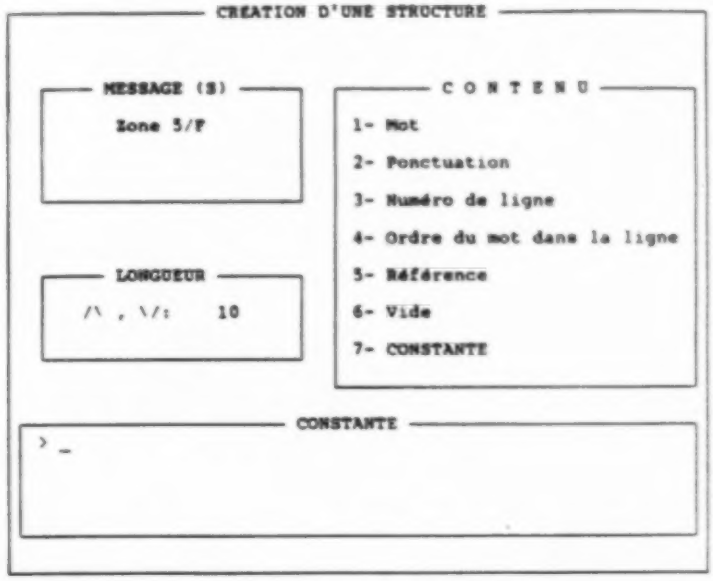
Avec la fenêtre *délimiteur droite* nous procédons exactement de la même façon qu'avec la fenêtre *délimiteur gauche*. La seule différence est que nous choisissons le caractère utilisé en guise de délimiteur droite pour la référence dans le texte. Choisissons le caractère) qui a la valeur ASCII 41.

FIGURE 15



Lorsque le délimiteur droite a été identifié, le module retourne au premier plan d'interrogation et demande la description de la zone suivante. Pour l'exemple, il s'agit de la cinquième zone. Pour cette zone, nous indiquerons que nous voulons y placer une constante. Le module fait apparaître une nouvelle fenêtre qui s'intitule *constante* (voir figure 16).

FIGURE 16



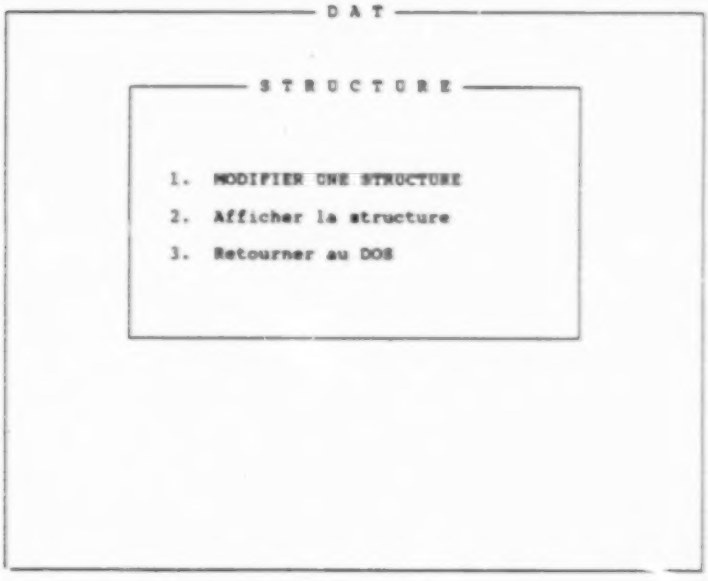
En faisant apparaître cette fenêtre, le module demande de faire la saisie de la constante à placer dans cette zone. Il suffit d'en faire la saisie, de la même façon qu'on le ferait avec un logiciel de traitement de texte. Pour l'exemple, la constante sera : *0000000000*.

Jusqu'à maintenant cinq zones ont été définies. La première contient un mot, la seconde un signe de ponctuation, la troisième une référence à position fixe, la quatrième une référence à position variable et la dernière une constante. Pour mettre fin à la séance de création d'une structure on appuie sur la touche **Esc** du clavier. Le module nous ramène donc au menu principal (voir figure 17).

3.3 L'affichage d'une structure

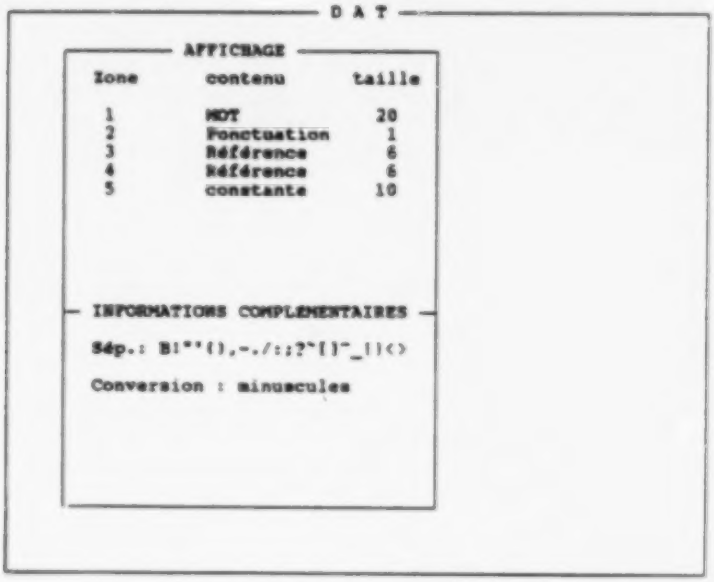
Comme on peut le voir, la première option du menu n'est plus *Création d'une nouvelle structure*, mais plutôt *Modifier une structure*. Puisque nous avons déjà créé une structure nous ne pouvons que la modifier ou la faire afficher (voir figure 18).

FIGURE 17



Pour l'affichage, il y a deux fenêtres. Dans la première, celle du haut, figure l'information qui a été obtenue avec le premier plan d'interrogation : le contenu et la taille des zones. Dans la seconde fenêtre apparaît l'information qui a été obtenue avec le second plan d'interrogation. Le contenu de la seconde fenêtre correspond à la zone dont le contenu est écrit en caractères majuscules. Dans l'exemple, le contenu de la seconde fenêtre correspond à la zone 1, qui contient le mot. Il est possible de voir l'information complémentaire qui est rattachée aux autres zones en utilisant les touches **flèches haut** et **flèches bas**. Par exemple, en choisissant la quatrième zone apparaissent les informations suivantes : le délimiteur gauche (et le délimiteur droite, qui est) (voir figure 19).

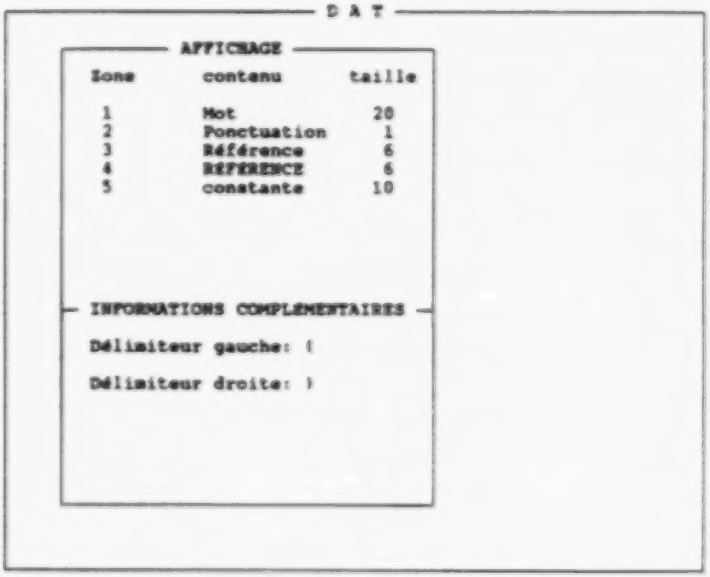
FIGURE 18



4. Le découpage

La séance d'interrogation de l'utilisateur est sans conteste l'étape la plus délicate du système de découpage. La qualité du résultat dépend directement de cette opération. Le processus de découpage s'active d'une façon très simple. Afin de procéder au découpage, il faut donner au système deux instructions.

FIGURE 19



La première instruction sert à indiquer au système le nom du fichier de directives qu'on désire utiliser pour le découpage. Supposons que le nom du fichier de directives qu'on désire utiliser est *lemme.str*. Il s'agit de donner l'instruction suivante : **DAT lemme.str/c**.

Lorsqu'on a indiqué le nom du fichier de directives on peut alors procéder au découpage. Supposons que le nom du fichier à découper soit *texte.txt* et que le résultat doive être placé dans un fichier du nom de *resultat.txt*. Il faut donner l'instruction suivante : **DAT texte.txt resultat.txt**. Le module de découpage se met alors en marche et le travail se fait à très grande vitesse.

5. Développement à venir

Certaines modifications seront apportées au système de découpage afin qu'il puisse lire un dictionnaire, dont la saisie a été faite à l'aide d'un logiciel de traitement de texte, et le découper de façon à permettre son automatisation.

Avec un fichier de directives appropriées, le système de découpage sera en mesure de distinguer la micro-structure de la macro-structure et de distinguer également tous les éléments constituant la micro-structure. Il constituera une fiche par entrée du dictionnaire. Dans chaque fiche il y aura une zone pour le mot et une zone par élément de la micro-structure.

Un dictionnaire ainsi découpé pourra être transmis à un logiciel de gestion de banque de données qui permettra un accès direct à chacun des éléments de la macro-structure et de la micro-structure.

UN MODULE LINGUISTIQUE POUR L'INFORMATIQUE GOUVERNEMENTALE DES ANNÉES 1990[1]

Richard Parent
Ministère des Communications
Gouvernement du Québec

Le rôle important que la langue française est appelée à jouer dans les systèmes informatiques de nos organisations constitue l'objet de cette communication. Nous allons d'abord préciser la vision prospective d'un module d'informatique linguistique de façon théorique, puis nous donnerons l'exemple de ce qui a été entrepris avec des moyens encore expérimentaux au sein de l'administration publique québécoise.

1.0 Prospective

Dans une perspective générale, la figure 1 illustre l'évolution des principaux usages de l'informatique dans une grande organisation nord-américaine typique :

FIGURE 1

Évolution de l'usage de l'informatique dans une grande organisation.

Années	Objet du traitement	Méthodes et techniques
1965	DONNÉES	- emmagasinage, classement, calcul
1980	INFORMATION	- réseaux, interfaces, transactions, SGBD[2], télématique, génie logiciel
1995	CONNAISSANCES	- systèmes-experts et traitement de la langue naturelle, moteurs d'inférences, analyseurs et dictionnaires.

Les perspectives à long terme évoquées ci-après s'inspirent des visions de chercheurs réputés (Sowa 1984; Johnson 1985 et Tenenbaum 1987). Johnson parle d'un «système de langage corporatif» qui sera basé sur le vocabulaire défini par son domaine particulier, capable d'accommoder diverses vues externes selon les besoins des usagers, mais ayant une structure conceptuelle définie via des ou-

[1]Communication présentée lors de la Deuxième Conférence des chefs d'État et de Gouvernement de pays ayant en commun l'usage du français, sous le thème : Industries de la langue, Québec, 5 septembre 1987.

[2]SGBD = Système de gestion de bases de données.

tils communs à l'ensemble de l'organisation. Pour sa part Tenenbaum, en se référant aux barrières actuelles de complexité des systèmes experts, parle du besoin de se donner des outils de développement plus puissants, en particulier des outils spécialisés pour certains domaines d'application avec des banques de connaissances spécialisées et d'autres banques plus générales. Pour ces dernières, il mentionne l'importance des concepts et des relations de temps, d'espace, de causalité comme outils utilisables de façon profitable dans le développement d'applications.

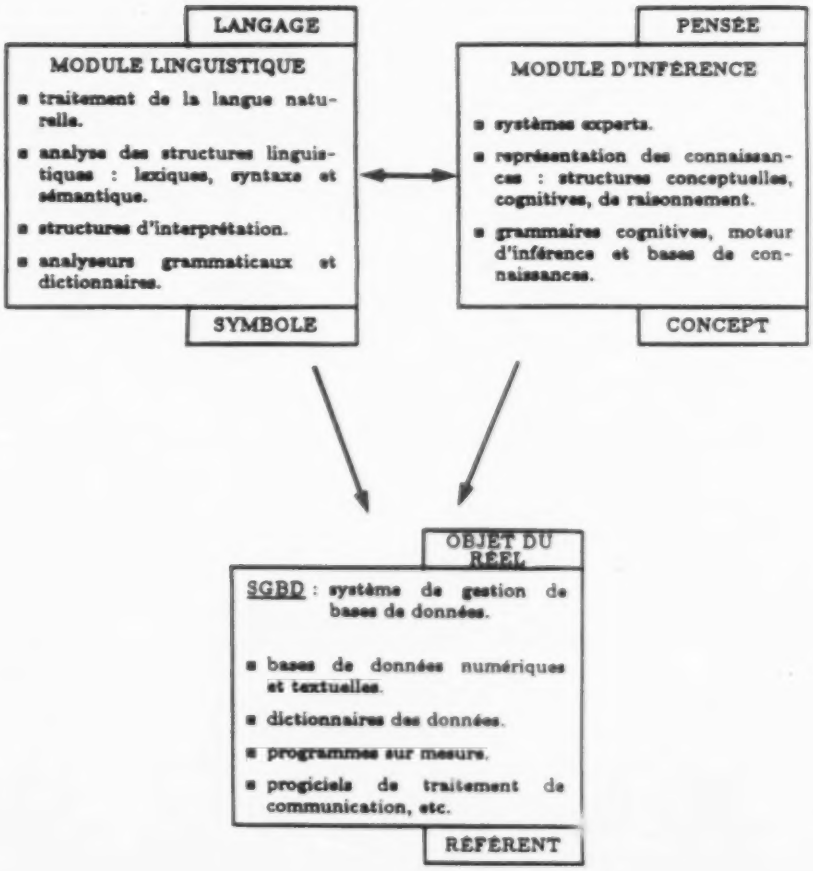
1.1 Place du module linguistique

La figure 2 présente la place éventuelle d'un module linguistique dans les systèmes informatisés de la fin du siècle. Le schéma compte trois grandes boîtes interreliées et interdépendantes : le module linguistique et le module d'inférence sont encore inexistants, sauf à l'état expérimental dans quelques grandes organisations comme **Bolt, Beranek & Newman, Boeing** et quelques autres.

Ainsi, l'ensemble des systèmes existants se retrouve dans la boîte **SGBD**, dans l'esprit de ce qui existe pour des systèmes multi-programmes, multi-usagers, de grand volume et avec des mécanismes de contrôle pour la sécurité, la mise en réseau, etc.

FIGURE 2

Place du module linguistique



Le module linguistique décrit les structures et les matériaux que la langue française (y compris les anglicismes) fournit à nos processus mentaux : lexique, syntaxe et sémantique sont combinés pour représenter l'information (données) dans le SGBD, ainsi que l'information (connaissances) dans le module d'inférence. La mise en correspondance des modules se fait grâce à un langage formel commun (maintenu invisible aux usagers grâce à l'interface pour un dialogue écrit en un français qui devra toujours subir certaines contraintes).

1.2 Composantes du module linguistique

Essayons maintenant de décrire un peu plus précisément ce que pourrait contenir un module linguistique. Trois types de composantes sont plausibles :

1- Un analyseur grammatical du français contenant des informations sur le lexique, la morphologie et la syntaxe (sorte de «moteur linguistique», par analogie au moteur d'inférence des systèmes experts).

2- Des dictionnaires de la langue française comprenant les informations et les connaissances lexicales mentionnées plus haut, mais aussi des informations, des connaissances et des structures d'interprétation sémantique. Il y aura au moins deux types de dictionnaires du français :

   a) le dictionnaire de *langue générale* (ex. : *Le Robert*) avec des définitions, des classifications et des hiérarchies conceptuelles, se prêtant à divers usages (sémantique, synonymie, associations, variations régionales);
   b) le dictionnaire de *langue spécialisée* : terminologies scientifiques et techniques, langues de spécialité, vocabulaires de domaines (ex. administratif).

   Ces divers dictionnaires auront un double usage : certaines suites de symboles sont destinées à être lues par l'ordinateur et d'autres sont destinées à l'affichage pour être lues par l'utilisateur.

3- Des outils de programmation pour la prise en compte du langage (administratif et autre) propre à une organisation. Cette composante sera ajoutée au module linguistique pour inclure le vocabulaire du domaine d'activité de l'organisation, les structures conceptuelles, les règles de fonctionnement, les normes administratives, les liens sémantiques avec les bases de données, etc.

1.3 Applications liées au module linguistique

Dans «Potentiel d'application de Déredec dans le contexte de la bureautique», Pierre Plante (1986) avait divisé les applications en deux grandes catégories, soit les situations d'écriture et les situations de lecture. La liste qui suit provient en grande partie de cette étude, mais quelques éléments ont été ajoutés, particulièrement en ce qui concerne les liens les plus directs avec les systèmes-experts.

À noter que la distinction entre lecture et écriture n'est pas toujours évidente. Quand on écrit, on se lit et se relit forcément, mais l'écriture est considérée comme dominante. Il faut voir la distinction comme étant relative, puisqu'il y a presque toujours un mélange des deux types d'activité dans les situations réelles.

Voici donc une liste représentative, mais non-exhaustive, d'applications potentielles du module linguistique à l'informatisation du travail de bureau. L'aide apportée à l'usager porterait sur :

1- *l'écriture :*

- correction orthographique et grammaticale;
- vérification de termes (uniformisation);
- recherche d'analogies (ex. le *Petit Robert* est un dictionnaire analogique;
- préparation de résumés;
- établissement d'un thésaurus (répertoire alphabétique des termes normalisés);
- indexation de documents;
- mesures de lisibilité, de redondance lexicale et remarques sur le style;
- vérification de cohérence de certaines structures sémantiques;
- rédaction d'un document «normé» (rapport, contrat, mémoire, compte-rendu);
- «complétion» d'un formulaire «intelligent» (capable de s'adapter grâce aux réponses antérieures de l'usager avec l'aide du module d'inférence);
- interfaces pour l'interrogation des bases de données de l'organisation et à l'extérieur, pour l'usage de progiciels statistiques et autres, pour certains systèmes de transaction dans les bureaux, pour les services télématiques grand public;
- éditeur pour la modélisation et la programmation de systèmes d'information et de systèmes experts au moyen de grammaires cognitives, d'ensembles terminologiques, de règlements et normes maison, d'outils pouvant être reliés au SGBD, etc.

2- la *lecture :*

- repérage dans les bases de données numériques et textuelles à partir d'éléments lexicaux ou de contraintes syntaxiques ou sémantiques;
- exploration et analyse de contenu d'un corpus donné de textes en vue, par exemple, d'en faire une synthèse, ou de les comparer entre eux, ou de relever les principaux propos sur divers thèmes et de les analyser;
- classification documentaire selon un thésaurus pré-établi ou selon des critères lexico-sémantiques mesurés à partir des textes mêmes.

2.0 Exemple de l'administration publique québécoise

La vision prospective des applications peut être prise comme une série d'objectifs à long terme en vue d'accroître la productivité dans l'administration publique. C'est ce qui inspire les actions du Bureau central d'informatique, en raison de son rôle de coordination dans le secteur public, qui est de favoriser l'apprentissage de nouvelles méthodes et l'application de nouvelles techniques dérivées des recherches en intelligence artificielle. L'attention a surtout porté sur les systèmes experts, mais les applications linguistiques font l'objet de travaux importants sur le point d'être amorcés. Dans les deux cas, les objectifs sont les mêmes :

- identifier les types et les zones d'application les plus prometteuses;
- contribuer à la conception de modèles et à la mise au point d'outils pour adapter ailleurs ces premiers modèles;
- viser à faciliter la combinaison entre des outils logiciels pour l'analyse linguistique et ceux pour la représentation des connaissances, et à l'interaction requise avec les systèmes informatiques existants.

La concrétisation de ces objectifs se fait par l'intermédiaire de projets-pilotes où l'on tente de réaliser des modèles généraux sous forme de progiciels répondant à des besoins spécifiques, mais combinables grâce à un système formel commun (Déredec). C'est ainsi qu'a été mis au point un

Progiciel de Génération de Système Expert (**PGSE**) qui fait l'objet d'un perfectionnement continu, principalement par contrat avec des chercheurs universitaires. L'intérêt de cet outil est de préfigurer le module d'inférence appelé à interagir étroitement avec le module linguistique.

Les projets relevant du module linguistique sont déjà importants :

- un progiciel de correction orthographique et grammaticale, utilisable en mode interactif, et conçu pour des élèves de 6e année (ministère de l'Éducation);
- un système d'aide au dépouillement terminologique (Office de la langue française);
- groupe d'étude interministériel dont les participants identifient un corpus de textes significatifs pour leur organisation en vue d'en mieux connaître le lexique et d'évaluer les possibilités d'application de modèles d'enrichissement sémantique par rapport au vocabulaire de leur domaine et à leur contexte en termes de système d'information. Le titre de ce projet juxtapose deux phases : «Projet lexical et sémantique de domaines». Trois choix importants ont guidé la conception du projet : le niveau de langage, les domaines d'application, les logiciels utilisés.

2.1 Choix du niveau de langage

Le choix des «zones» de langage à étudier en vue de les représenter en priorité dans l'administration publique relève dans un premier temps des niveaux de contenu. Ce n'est pas dans l'administration publique que doit être effectué le travail sur le vocabulaire de base et général du français. L'évaluation doit porter sur des domaines bien circonscrits, à partir de critères semblables à ceux qu'on utilise pour juger de la faisabilité d'un système expert.

Il est évident qu'on peut se perdre dans l'immensité du langage si on cherche à tout couvrir. Comme pour les systèmes-experts, on doit se limiter à des vocabulaires restreints (quelques centaines ou quelques milliers de mots) dont la représentation automatisée en dictionnaire évolutif permet d'envisager des applications rentables. Une approche étapiste ainsi que des résultats «intermédiaires» utilisables constituent des moyens pour explorer concrètement des hypothèses sur le module linguistique : on pourra alors s'intéresser notamment à la taille des vocabulaires, à l'importance du chevauchement des domaines, aux aspects cognitifs des structures sémantiques d'un domaine, etc.

2.2 Choix des domaines d'application

Le choix des domaines d'application qui suit a été fait progressivement au cours des derniers mois grâce à des échanges avec des professionnels de la fonction publique intéressés aux systèmes experts. Le résultat provisoire de ce processus de consultation a permis d'identifier six domaines d'application qui apparaissent intéressants pour différentes raisons.

1- L'informatique : analyser des dictionnaires de données des SGBD relationnels existant dans la fonction publique et d'autres modèles de données en vue d'aider à la modélisation et à l'administration conceptuelle des données.

2- La gestion documentaire : analyser les plans de classification d'organismes du secteur public et du ministère en vue d'aider à la classification et au repérage des documents.

3- Les normes administratives : analyser les règlements, les normes, les procédures et formulaires dans un secteur intéressant le Conseil du Trésor.

4- Le calcul et l'analyse statistique : analyser la terminologie du domaine de la statistique et des éléments de données dans deux ministères, en vue de favoriser l'accès aux outils statistiques ainsi qu'aux données emmagasinées dans leurs banque de données.

5- L'environnement : analyser le vocabulaire administratif et règlementaire de l'environnement en usage dans ce ministère (règlements, orientations, politiques) ainsi que le vocabulaire relatif à l'administration d'un programme hautement scientifique (Système d'aide à la gestion des évaluations environnementales.

6- Les municipalités : analyser le vocabulaire des Systèmes d'information urbains à références spatiales ou celui d'un guide d'information sur les municipalités préparé récemment.

Pour les trois premiers domaines, (informatique, gestion documentaire et normes administratives), on peut viser le développement d'outils horizontaux qui seraient adaptés localement tout en conservant une cohérence d'ensemble : ce sont d'importants champs d'application pour l'administration publique.

Le quatrième domaine (calcul et analyse statistique) est en partie une affaire de langage d'une spécialité scientifique, mais il constitue aussi un accès aux bases de données d'une organisation en des termes familiers pour «quelqu'un du domaine».

Les deux derniers domaines (environnement et municipalités) constituent des secteurs d'activités déterminés de l'administration publique; ils ont été choisis parce qu'ils présentaient des possibilités d'application intéressantes en bureautique et en télématique. Ils sont toutefois différents : avec l'environnement, l'accent est mis sur une terminologie scientifique et sur la bureautique (systèmes experts) alors qu'avec les municipalités, l'accent est semblable pour les S.I.U.R.S., mais la terminologie est plus vulgarisante dans le guide d'information.

Différents principes ont donc inspiré le choix des domaines d'application. Le tout constitue un premier échantillon exploratoire de la langue de l'administration publique.

2.3 Choix des logiciels utilisés

Déredec est choisi en tant qu'outil formel, en tant que système existant de description linguistique du français et parce qu'il est compatible avec les structures de représentation requises par les systèmes experts. Cela n'exclut absolument pas sa combinaison éventuelle avec d'autres produits existants. Le choix de **Déredec** comme outil principal de support se comprend facilement eu égard au développement en cours de **Déredec-EXPERT**, ainsi que des projets qui seront bientôt entrepris sur l'Écritoire (avec le ministère de l'Éducation) et sur un système d'aide au dépouillement terminologique (avec l'Office de la langue française).

Conclusion

Ce qu'il faut comprendre pour croire à cette démarche, c'est que la langue française est en train de devenir un objet d'information. Il faut prévoir une croissance rapide du marché au cours des années 1990 et ne pas investir là où des efforts dans des domaines ou sur des aspects trop universels seraient inutiles compte tenu de produits commerciaux existants. Une organisation doit plutôt investir dans l'étude de ses particularités, de façon à devenir complémentaire aux produits qui existent déjà ou qui sont susceptibles d'apparaître sur le marché : c'est l'esprit du «projet lexical et sémantique de domaines». Cette démarche permettra en même temps de commencer à explorer les contraintes et les capacités d'un éventuel «module linguistique» en interaction avec un «module d'inférence» et les SGBD accessibles sur réseau ou localement.

BIBLIOGRAPHIE

JOHNSON (Tim). *Natural Language Processing: Commercial Applications*. OVUM. Londres. 1985.

Le Petit Robert. 1978. Nouvelle édition.

Manuel d'utilisation Déredec.

PLANTE (Pierre), DUCHASTEL (Jules) & (Lorne) BOUCHARD. *Potentiel d'applications de Déredec dans le contexte de la bureautique*. Rapport d'étude produit au Centre d'ATO, UQAM et distribué par le Ministère des Communications, Direction générale des technologies de l'information. 1986.

SOWA (John F.) *Conceptual Structure. Information Processing in Mind and Machine*. Addison-Wesley, Reading, MA, 1984.

TENENBAUM (Jay M.). "Artificial Intelligence: closing the gap between expectations and reality". The Third Conference on Artificial Intelligence Applications. Orlando, E.U. 23 au 27 février 1987.